Université de la Méditerranée

Habilitation à diriger les recherches

présentée et soutenue publiquement par

Claude Manté

le 06 Juin 2008

De la nature de quelques types de chroniques environnementales. Conséquences pour leur traitement statistique.

Commission d'examen :

Pr. Avner Bar-Hen	Université Université Paris V	Rapporteur
Pr. Philippe Besse	Institut de Mathématiques de Toulouse	Rapporteur
Pr. Claude Deniau	Université de la Méditerranée	Rapporteur
Pr. Yvan Lagadeuc	Université Rennes I	Examinateur
Pr. Robert Sabatier	Université Montpellier I	Examinateur
Pr. Jean-Pierre Durbec	Université de la Méditerranée	Directeur de Recherche

Remerciements

Je souhaite tout d'abord remercier Philippe Besse, qui a immédiatement accepté d'être rapporteur de ce travail, ainsi qu'Yvan Lagadeuc et Robert Sabatier qui m'ont fait l'amitié de participer au jury, en tant qu'examinateurs.

Que serait ce mémoire devenu sans les amicales pressions de Claude Deniau, Jean-Pierre Durbec et Avner Bar-Hen ? Un avorton, un embryon desséché ! Ce travail est en partie le fruit de leurs encouragements, et je les remercie d'en être également rapporteurs.

Ma reconnaissance va également aux lecteurs attentifs de nombreuses versions provisoires ou partielles du manuscrit : M. Brémond, B. Mérigot, J.C. Poggiale, A. F. Yao, et d'autres...

Si je dois beaucoup aux discussions et travaux avec mes "complices", statisticiens ou mathématiciens appliqués (J.P. Durbec, J.C. Gaertner, D. Nerini, J.C. Poggiale, A.F. Yao), j'ai aussi une dette envers nombre de collègues du COM, qui m'ont donné accès à leurs données et à leurs compétences : F. Carlotti, P. David, C. Degiovanni, F. Gilbert, G. Gregori, G. Miralles, B. Thomassin, entre autres. Je remercie au même titre d'autres océanologues d'autres laboratoires, rencontrés dans cadre (amical, incitatif et productif) du Programme National D'Océanographie Côtière, prolongé par le Programme National Environnement Côtier (PNEC) : les benthologues J.C. Dauvin et B. Elkaim, ainsi que quelques collègues du Laboratoire d'Océanographie de Villefranche : G. Gorsky, F. Ibanez et A. Sciandra. Je remercie également le Groupement d'Intérêt Public pour la Réhabilitation de l'Etang de Berre, qui a pris en charge les campagnes sédimentologiques de 1992 et 1997 que j'ai utilisées.

Je n'oublie pas non plus que lorsque j'étais à Caen, dans une vie antérieure, j'ai été incité à rédiger mes premières publications par mes collègues J.P. Coutard, J.C. Ozouf et B. Francou, que je remercie. Notre travail sur les séries chronologiques et les courbes granulométriques est à l'origine des chapitres 1 et 3 de ce document.

Enfin, je terminerai par une pensée spéciale pour celui qui fut mon Directeur de Thèse, Guy Der Megreditchian. Les quelques années passées à son contact à la Météorologie Nationale m'ont fait prendre conscience de mon goût pour la recherche (et de mon aversion pour toute autre forme de travail).

Table des matières

Introduction

> *Tous les grands progrès théoriques, à mon avis, proviennent de la capacité des inventeurs à "se mettre dans la peau des choses", pour pouvoir s'identifier par empathie à n'importe quelle entité du monde extérieur. Et cette espèce d'identification transforme un phénomène objectif en une sorte d'expérience concrète et mentale.* [**Th93**, p. 92]

L'essentiel de ce qui va être présenté ici est le résultat de collaborations avec des chercheurs ou des doctorants d'autres disciplines que la mienne (géographes, océanologues, géologues, biologistes, écologues). Le lecteur de ce document y trouvera donc exposées avec plus ou moins de détails, les analyses de données environnementales très diverses : météorologiques, physico-chimiques, acoustiques, images, profils d'éboulis, relevés benthiques, courbes granulométriques, profils d'électrophorèse, chroniques d'énergie cinétique turbulente.

Dans cet inventaire à la Prévert, où sont les ratons laveurs, où est la régularité ? Elle réside d'une part dans l'approche exploratoire généralement adoptée, d'autre part dans le fait que dans les Sciences de l'Environnement, les observations sont pratiquement toujours indicées par une variable continue - le plus souvent le temps. Cela m'a souvent conduit à utiliser les méthodes de l'Analyse des Données Fonctionnelles, à la mode aujourd'hui [**RS05, FV06**], mais dont l'origine remonte à Deville [**De74**]. D'autre part, et c'est probablement une originalité de ce travail de **Statistique Appliquée**, je me suis efforcé de prendre en compte à chaque fois la nature des données dans le traitement, en considérant en quelque sorte leur paradigme. J'entends par là leur nature physique et l'objectif visé par l'interlocuteur qui les a collectées (écologue, géologue, biologiste, géographe, physicien, ...), en plus de leurs caractéristiques mathématiques (positivité, périodicité, monotonie, intermittence, *etc*). Il n'est en effet pas raisonnable (et si ennuyeux !) de passer à la même "moulinette" des observations si diverses, alors que, comme le font remarquer Ramsay & Silverman, une approche ouverte est plus payante :

> *... with each new set of functional data, we have discoverd challenges and invitations to develop new methods. Statistics shows its finest aspects when exciting data find existing statistical technology not entirely satisfactory.* [**RS05**, p. 384]

Exemples

Il est fréquent que l'instant d'origine de séries chronologiques soit totalement arbitraire. Ce fut le cas par exemple d'enregistrements **acoustiques** où l'on souhaitait mettre en évidence un éventuel signal noyé dans du bruit, puis l'identifier (Chapitre 1). Pour analyser ce genre de données, j'ai proposé une méthode exploratoire **invariante par translation,** ou, plus généralement, sous l'action d'un **groupe d'isométries**. Cette approche conduit à représenter chaque signal dans un espace abstrait (une variété quotient). L'analyse consiste à obtenir une image euclidienne de l'ensemble des "signaux" situés sur cette variété.

Dans le cas de séries chronologiques de relevés benthiques (Chapitre 2), où plusieurs centaines d'espèces sont comptées à chaque observation, j'ai proposé des méthodes permettant de traiter équitablement espèces rares ou communes. Pour ce, j'ai introduit plusieurs définitions de la **rareté**, correspondant à diverses approches (locale, globale), et élaboré plusieurs méthodes de sélection de variables/espèces, basées sur la rareté. Dans ce chapitre, la démarche est donc simultanément exploratoire et décisionnelle, et vise à répondre à des problèmes statistiques (procédures de sélection de variables) en formalisant une notion **écologique** (la rareté).

Par nature, les courbes granulométriques sont de leur côté assimilables à des **mesures de probabilité**, caractéristique que nous avons exploitée pour traiter les données issues de campagnes sédimentologiques (Chapitre 3). La représentation de ces mesures par des densités, très utilisée par les géologues, dépend simultanément de l'échelle choisie et d'une mesure (dite de référence) associée à la structure Hilbertienne choisie pour l'analyse. Cette mesure peut être arbitraire (uniforme, par exemple), mais peut aussi avoir une signification **géologique** (sédiment de référence, fonction de transport sédimentaire). Le cadre **exploratoire** proposé généralise les méthodes usuelles (Analyse en Composantes Principales et Analyse des Correspondances), en amalgamant le point de vue exploratoire avec celui de l'**Estimation de la Densité**.

Précisions supplémentaires

Afin de ne pas lasser le lecteur en rédigeant une fastidieuse compilation, j'ai parfois illustré mon propos par d'autre figures ou même d'autres données que celles associées aux articles joints (Chapitres 1 et 3). J'ai également introduit quelques "nouveautés" non publiées, soit parce que leur absence nuisait à la complétude de l'exposé (Section 3.1), soit parce qu'il s'agit de travaux en cours (Chapitre 4).

Enfin, pour ce qui concerne la bibliographie, les publications dont je suis co-auteur ont été placées en tête, et numérotées dans l'ordre chronologique ; dans cette liste, les publications jointes au document final sont "étoilées", comme par exemple [**P8***].

CHAPITRE 1

Aspects géométriques de l'Analyse en Composantes Principales de courbes : applications en acoustique sous-marine et en imagerie

Sommaire

De nombreuses données environnementales présentent une structure périodique (quotidienne, saisonnière, annuelle) marquée, tout en étant a priori non-stationnaires (même à l'ordre deux), voire non-linéaires. Par exemple, la moyenne et la variance d'une chronique de températures ne sont généralement pas les mêmes à midi ou à minuit, d'une saison à l'autre, *etc.* De plus, les trous généralement présents dans ces chroniques rendent difficile l'utilisation rigoureuse des méthodes classiques de traitement des séries chronologiques, même lorsqu'elles sont échantillonnées à pas fixe. Nous avons proposé des variantes de l'Analyse en Composantes Principales (ACP), conçues pour tenir compte de cette structure temporelle. Ces méthodes appartiennent au champ de ce que l'on appelle aujourd'hui l'Analyse des Données Fonctionnelles [**RS05, FV06**]. Cependant, alors que les auteurs du domaine mettent plutôt l'accent sur la régularité des courbes échantillonnées, nous nous sommes surtout intéressés au côté géométrique du problème, développé dans la Section 2, ainsi qu'aux capacités de filtrage de l'ACP, décrites ci-dessous.

1. L'ACP dans l'espace des fréquences

C'est une variante assez élémentaire : le signal (éventuellement multivarié) est d'abord découpé en ***blocs*** de longueur fixe $\boldsymbol{T}$, puis recodé par transformée de Fourier discrète et soumis à une ACP. La transformation de Fourier intervient de plusieurs manières dans cette méthode :

(1) elle permet de diminuer considérablement le nombre de variables à traiter, le signal étant généralement bien reconstitué avec un petit nombre d'harmoniques [**P1, P3, P5, P7**],

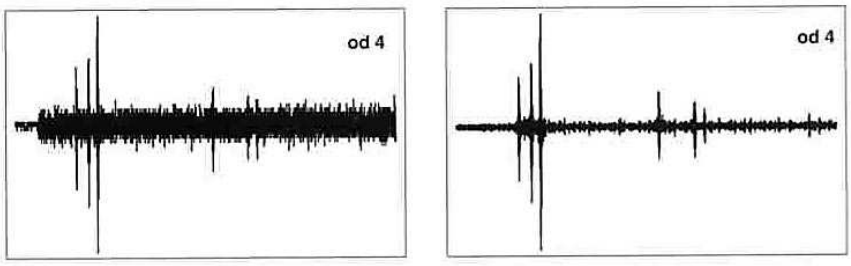
(2) elle permet si nécessaire d'équilibrer le rôle dans l'ACP de différents groupes de fréquences [**P1, P3, P5**], en utilisant par exemple la stratégie de pondération de l'Analyse Factorielle Multiple (***AFM***) [**EP90**],

(3) le fait qu'elle transforme les opérateurs différentiels linéaires à coefficients constants en polynômes permet de construire des "filtres" dont l'effet est de lisser et/ou dériver le signal [**P5**].

Dans le cas d'un signal univarié, chaque bloc est ainsi considéré comme le résultat de l'échantillonnage d'une fonction de l'espace de Hilbert $C(T)$, complété de l'espace des fonctions continues périodiques sur $[0, T]$. Il est ensuite projeté sur le sous-espace H de $C(T)$ **engendré par les harmoniques retenues**. Dans le cas multivarié (par exemple mesures de température à $K = 3$ profondeurs dans [**P1**]), les blocs appartiennent au sous-espace de dimension finie $\mathbb{H} := H_1 \times \cdots \times H_K$

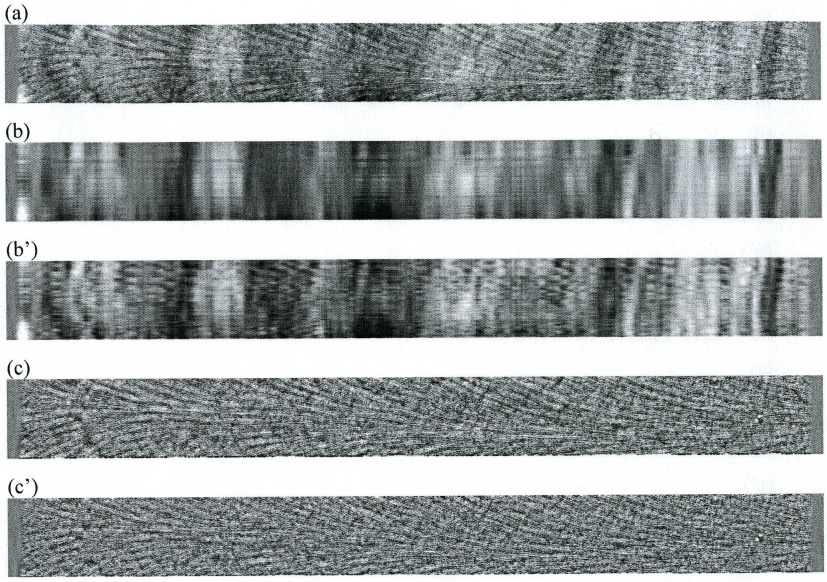
de l'espace produit $C(T)^K$ construit de la même manière (les espaces H_i sont indicés, car on ne retient pas forcément les mêmes harmoniques pour toutes les variables). On réalise ensuite l'ACP non normée des blocs dans $\mathbb{H}$. Nous avons analysé de cette façon des données météorologiques [**P1**], des images [**P5, P3**] et des signaux acoustiques [**P7**].

FIG. 1.1. ***Un signal acoustique. A gauche : l'enregistrement original ; à droite : le même, filtré par ACP (5 CPs).***



Un des intérêts de cette méthode est que la reconstitution des données à partir des premières composantes de l'ACP est une excellente méthode de filtrage. Ce fait est illustré par un premier graphique (Fig. 1.1), extrait de [**P7**]. Cette figure représente un enregistrement d'un signal acoustique rétrodiffusé par des bulles d'air présentes dans une colonne d'eau. Nous représentons les données brutes, puis filtrées par sélection des harmoniques et reconstitution par ACP. Les bulles insonifiées étaient calibrées, de rayon $345\mu m$, et l'enregistrement était de longueur 10^4 ; l'axe horizontal est la profondeur.

FIG. 1.2. ***Une coupe : son image brute (a), puis reconstituée avec deux (b) et sept composantes (b') ; les images résiduelles sont (c et c').***



La Figure 1.2, extraite de la thèse de P. Chevrot [**CH95**], est une autre illustration des performances de l'ACP dans l'espace des fréquences pour le filtrage de signaux. Il s'agit d'images de la densité du squelette d'une coupe de colonie corallienne. Les variations de densité (en niveaux de gris) sont liées à la croissance de ces organismes, qui dépend des forçages environnementaux. Il était donc primordial de mettre en évidence sur l'image 1.2(a) les "bandes de croissance" de la colonie, afin de remonter aux variations climatiques du site de prélèvement. Ici, ces bandes devaient être approximativement parallèles à l'axe des ordonnées de la figure, en raison de la structure locale du madrépore. Afin d'optimiser le contraste de l'image reconstituée, nous avons couplé l'ACP de cette image avec celle de son gradient horizontal. Pour ce, nous avons réalisé

l'AFM des deux groupes de variables représentant d'une part chaque ligne de l'image brute (appartenant à l'espace H), d'autre part la dérivée lissée de cette fonction (une ligne de l'image du gradient, appartenant à un espace noté ∂H). C'est une ACP dans $H \times \partial H$, qui affecte à chacun des deux espaces supplémentaires un poids optimisant la ressemblance entre les deux nuages de lignes associés [**EP90**]. Les deux premières composantes de l'AFM étant étroitement associées aux deux images (composantes "consensuelles"), les images associés devaient présenter un bon contraste horizontal, et correspondre à la structure recherchée (fort gradient horizontal). On peut constater *de visu* sur la figure 1.2 que la décomposition $(a) = (b) + (c)$ sépare bien la mésostructure de la microstructure en éventail (c) du madrépore. Si l'on incorpore des composantes non-consensuelles (de rang ≥ 3) dans la reconstitution, comme dans $(a) = (b') + (c')$, la décomposition n'est plus satisfaisante. En effet, la méso et la microstructure sont alors moins bien séparées, et les bandes de croissances moins nettes. Pour plus de détails, le lecteur se reportera à [**P5, P3, CH95**].

On peut cependant bien mieux exploiter les propriétés de la transformée de Fourier en partant d'un défaut de l'ACP dans l'espace des fréquences. En analysant des relevés de température [**P1**], nous avons pu constater que les fonctions propres (orthogonales) associées aux composantes principales successives peuvent en réalité être très semblables, dès lors qu'on les décale dans le temps. Or, l'origine temporelle d'un bloc de données n'a généralement pas de signification ! Cela m'a amené à proposer la méthode ci-dessous, qui m'a été inspirée par un article concernant l'usage des lois physiques de conservation pour classer des particules élémentaires d'après leur trajectoire [**LM78**].

2. L'analyse factorielle dans des espaces d'orbites

> *... on n'échappe pas à la nécessité de considérer des entités abstraites dans l'organisation de la réalité.* [**Th93**, p. 100]

Rappelons tout d'abord que le fondement même de la géométrie est la donnée d'un espace $\mathbb{H}$ et d'un groupe G de transformations opérant sur celui-ci, et que "les propriétés géométriques sont caractérisées par leur invariance relativement aux transformations du groupe" . [**Kl1872**].

Le groupe G correspondra ici à l'ensemble des symétries que l'on s'autorise pour identifier deux blocs, chacun d'entre eux étant assimilé à son orbite (***G-bloc***) appartenant à l'espace quotient $\mathbb{H}/G$. La ***distance orbitale*** entre deux G-blocs $\dot{B}_1$ et $\dot{B}_2$ est :

$$\mathcal{D}_{\mathrm{G}}(\dot{B}_1, \dot{B}_2) := \min_{g \in G} D(B_1, g \circ B_2), \tag{1.1}$$

où $D(B_1, B_2)$ désigne la distance (ici euclidienne) entre blocs. La méthode consiste à analyser la table des distances entre G-blocs, de manière à en produire une image euclidienne [**CP76**]. Elle a l'avantage de concentrer le nuage des blocs, en éliminant la forme parasite de variance liée aux symétries admises.

2.1. Invariance par translation (cas univarié).

Le groupe de symétries le plus naturel pour l'étude de séries chronologiques est celui des translations, noté Δ. Si, comme c'est ici le cas, les mesures sont échantillonnées avec un pas fixe τ, Δ est cyclique d'ordre $\mathcal{O}(\Delta) = T/\tau$ et opère sur l'ensemble des fonctions **périodiques** sur $[0, T]$. L'usage de la transformée de Fourier va donc de soi dans le cadre géométrique choisi.

Le générateur δ_1 de Δ est la translation de pas τ, associée dans l'espace des coefficients de Fourier H à la multiplication par $e^{2i\pi T/N}$. On choisit de préférence pour longueur des blocs un dyadique 2^p , de sorte que la transformée de Fourier rapide (FFT) soit utilisable. Par construction, la méthode hérite des avantages {1, 2, 3} de la transformée de Fourier.

Afin de réduire le temps de calcul de la distance (1.1), on peut faire agir un sous-groupe pertinent de Δ. Par exemple, dans [**P1**], le pas d'échantillonnage des températures était de deux heures, et nous avions choisi $\mathcal{O}(\Delta) = 2^7$, soit $T \approx 10.7$ jours. Il était donc nécessaire de faire opérer 128 fois δ_1 pour comparer deux G-blocs. Les données étant caractérisées par un cycle quotidien marqué, nous avons diminué le volume de calcul en ne faisant agir que le sous-groupe engendré par δ_4, la translation de 8H. C'est le seul sous-groupe strict de Δ engendrant la translation de

24H, il est d'ordre 32. Il suffisait alors de faire opérer 32 fois δ_4 pour avoir une approximation correcte de chaque interdistance, avec quatre fois moins de calculs.

REMARQUE. Chaque bloc est associé à un vecteur de $H = \mathbb{R}^q$, où q désigne le nombre de coefficients de Fourier conservés. Si l'on s'intéresse uniquement à la forme des courbes, il faut les normer en travaillant sur la sphère unité de l'espace des coefficients. Les transformations considérées étant des isométries, l'orbite d'un bloc normé appartient aussi à la sphère, et la méthode s'applique [**P7**].

2.2. Extensions du groupe de symétries (cas univarié).

Il est possible de construire de plus gros groupes de symétries, en considérant le produit de Δ avec d'autres groupes pertinents. Par exemple, dans [**P7, P1**], nous avons utilisé le groupe des antipodies, $\mathcal{A} := \{-1, +1\}$. En effet, un des problèmes abordés dans [**P7**] était de localiser les blocs contenant l'écho d'une ou plusieurs bulles d'air, les autres blocs ne contenant que du bruit blanc. Or, il est **physiquement** impossible d'avoir dans de tels enregistrements $\dot{B}_1 \approx -\dot{B}_2$, sauf si les deux blocs ne contiennent que du bruit (il n'existe pas d"'anti-bulle"). Nous avons donc utilisé le groupe $\Gamma := \Delta \times \mathcal{A}$, qui identifie les blocs B, $\delta_m \circ B$ et $-\delta_k \circ B$, avec la relation suivante entre les distances orbitales :

$$\mathcal{D}_\Delta(\dot{B}_1, \dot{B}_2) \geq \mathcal{D}_\Gamma(\dot{B}_1, \dot{B}_2).$$

Dans notre cas, l'inégalité ne peut être stricte que pour des blocs sans intérêt, puisque $\mathcal{A}$ ne peut agir concrètement que sur le bruit. Par conséquent, la substitution de Γ à Δ ne peut rapprocher que les blocs parasites, permettant une meilleure discrimination des blocs d'intérêt (voir les figures 1.3 et 1.4).

2.3. Le cas des données multivariées [P1].

Lorsque le signal est multivarié, les blocs appartiennent à $C(T)^K$, et le problème du synchronisme des K variables se pose. Si l'on considère qu'elles sont asynchrones, le même groupe G agit **séparément** sur les K courbes, et le groupe de symétrie est le produit G^K. La distance orbitale "naturelle" entre G-blocs est alors :

$$\mathcal{D}_G^1(\dot{B}_1, \dot{B}_2) := \sqrt{\sum_{i=1}^{K} \mathcal{D}_{\mathrm{G}}(\dot{B}_1^i, \dot{B}_2^i)^2}.$$

Si, par contre, on impose la synchronie des variables, le groupe de symétrie est un petit sous-groupe de G^K, sa "diagonale" $\{(g, \ldots, g) : g \in G\}$, et on obtient la distance orbitale :

$$\mathcal{D}_G^2(\dot{B}_1, \dot{B}_2) := \sqrt{\min_{g \in G} \sum_{i=1}^{K} D(B_1^i, g \circ B_2^i)^2}.$$

2.4. Digression géométrique .

Nous avons montré [**P1**] que $\mathbb{H}/G$ (débarrassé de l'improbable ensemble de ses singularités) est une variété quotient à $\mathcal{O}(G)$ feuillets, et que la distance orbitale entre G-blocs coïncide avec la distance géodésique sur cette variété. L'analyse du tableau des interdistances orbitales devrait ensuite fournir une image euclidienne fidèle du nuage de G-blocs, mais il faut tenir compte du fait que l'on rencontre généralement des valeurs propres négatives, car la géométrie de $\mathbb{H}/G$ est non-euclidienne.

Nous le montrons ci-dessous dans le cas d'un groupe de translations Δ de générateur δ. Considérons trois blocs : B_0 , B_1 et B_2. Il existe une paire de translations telles que l'on ait, pour $i = 1, 2$:

$$\mathcal{D}_\Delta(\dot{B}_0, \dot{B}_i) = D(B_0, \delta^{n_i} \circ B_i).$$

S'il existait une image euclidienne exacte de ces trois points, leur représentation serait obtenue avec B_0, $\delta^{n_1} \circ B_1$ et $\delta^{n_2} \circ B_2$ sur un même feuillet de H/Δ. Mais pour que cela soit possible, il faudrait que les translations soient compatibles, c'est-à-dire que l'on ait :

$$\mathcal{D}_\Delta(\dot{B}_1, \dot{B}_2) = D(\delta^{n_1} \circ B_1, \delta^{n_2} \circ B_2) = D(B_1, \delta^{n_2 - n_1} \circ B_2). \tag{1.2}$$

FIG. 1.3. ***Premier plan de l'ACP dans l'espace des fréquences.***

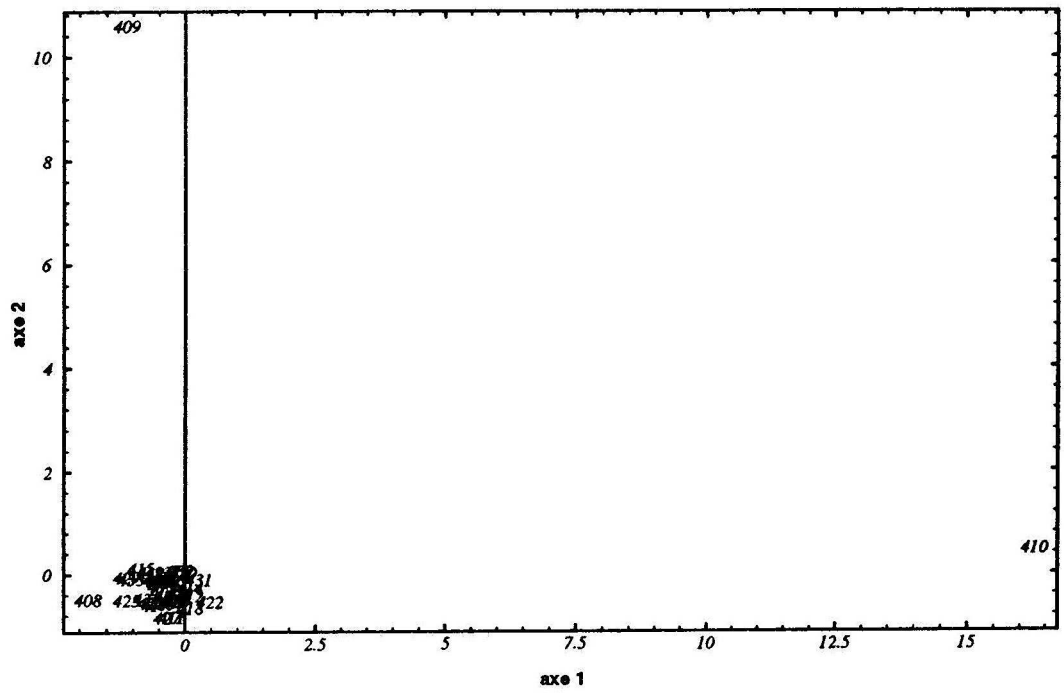


Figure 28 : A.C.P. sur od 4 -Plan factoriel 1/2-
Valeurs propres : vp1=7.359 ; vp2=3.108
Inertie : axe 1→43.65% ; axe 2→18.44%

Or, il n'y a aucune raison pour que cette égalité soit vraie : la distance orbitale peut être atteinte pour un décalage $n_{1;2} \neq n_2 - n_1$. Par conséquent, il n'existe pas en général d'image euclidienne parfaite de ce genre de données. Cependant, dans tous les cas rencontrés, le module de la plus petite valeur propre étant très petit, la représentation obtenue était valide.

2.5. Application en acoustique sous-marine.

Nous illustrons l'intérêt des distances orbitales par deux figures extraites de la thèse d'Ali Khelil [**AK95**]. Ces figures sont associées à l'une de ses expériences, dont l'enregistrement a été représenté sur la figure 1.1. L'un de nos buts était de localiser dans ce signal la présence d'échos correspondant à l'énergie rétrodiffusée par un nombre inconnu de bulles calibrées. Il est à noter que les bulles situées au fond de la colonne d'eau sont très difficiles à détecter, car l'énergie sonore diminue exponentiellement avec la profondeur (axe horizontal de la figure 1.1).

Nous voyons sur la figure 1.3 que l'ACP dans l'espace des fréquences permet de détecter des bulles dans les blocs 09 et 10, et peut être 08. Les mêmes blocs sont projetés sur la figure 1.4 , mais cette fois à partir de leurs interdistances sur H/Γ. L'utilisation de cette distance permet de régler le cas du bloc 08, qui contient bien une bulle, et d'en détecter d'autres, dans les blocs 22, 25 et 26. Notons aussi que les valeurs propres et les taux d'inertie sont nettement plus importants dans le dernier cas. L'existence de valeurs propres négatives pourrait tempérer notre enthousiasme, mais elles sont en valeur absolue très petites [**AK95**]. D'autres illustrations se trouvent dans [**AK95, P7**].

2.6. Quelques travaux apparentés.

Dans leur ouvrage, Ramsay et Silverman [**RS05**] traitent aussi du problème de la mise en coïncidence de courbes. Ils utilisent des décalages, mais aussi des déformations plus élaborées ("warping") de l'axe des abscisses (le temps, généralement) pour superposer les courbes. La deuxième option paraît trop complexe pour être incorporée dans une ACP : à chaque paire de

FIG. 1.4. ***Premier plan de l'analyse de la table d'interdistances entre les Γ-blocs de od4.***

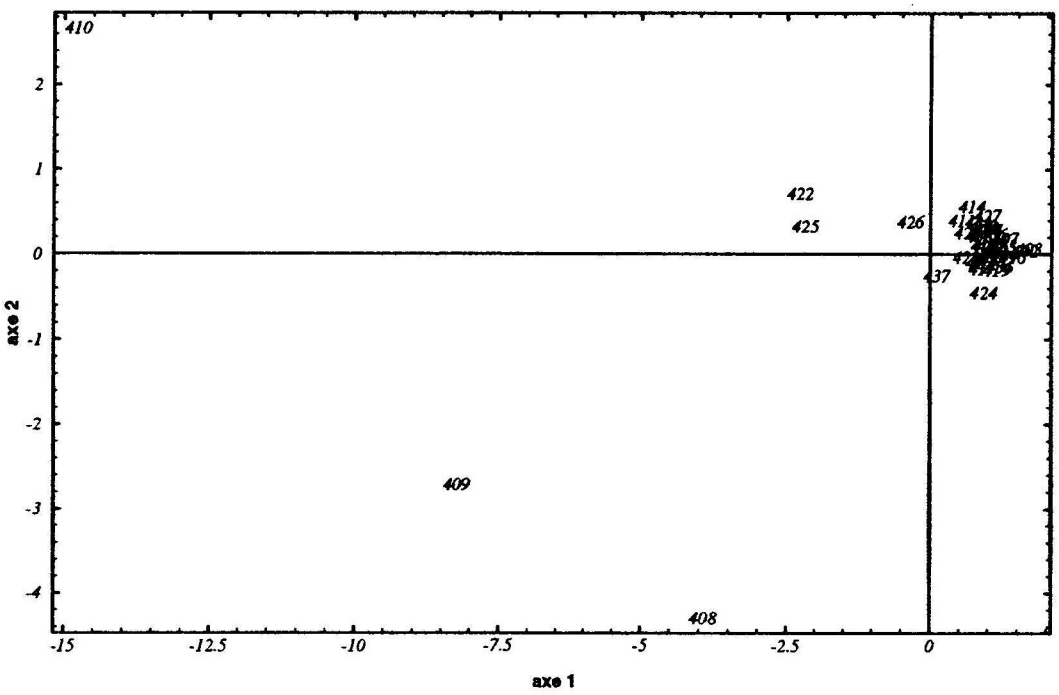


Figure 34 : A.F.T.D. sur od 4 ($D_{or}2$) -Plan factoriel 1/2-
Valeurs propres : vp1=9.09 ; vp2=0.93
Inertie : axe 1→ 69.30% ; axe 2→ 7.12%

courbes serait associée une troisième courbe (à estimer) mettant en correspondance les deux temps propres. Par contre, ils incorporent des décalages dans l'ACP de courbes en essayant de toutes les aligner au préalable sur leur moyenne empirique. Cette méthode se justifie pour des courbes très semblables, où la relation (1.2) peut être vérifiée avec une bonne approximation, mais nous avons vu dans la Section 2.4 qu'il n'existe pas en général de représentation euclidienne exacte d'un ensemble de courbes décalées.

Dans la mesure où ce travail faisait simultanément appel à l'analyse de Fourier classique (associée à la représentation du groupe des translations) et aux variétés quotient, il s'apparente plutôt à d'autres recherches. Je citerai d'un côté la "Théorie des formes" [**Ke89, LK00**, par exemple] dont l'approche procustéenne est proche de la notre. Cependant, ces auteurs analysent des nuages de points sans référence à une structure fonctionnelle particulière, et font intervenir un groupe de symétrie beaucoup plus grand que ceux utilisés ici (translations, rotations et homotéties). Pour ce qui concerne l'aspect fonctionnel des données, il existe une autre approche, impliquant l'analyse harmonique abstraite, c'est-à-dire à l'analyse de Fourier des groupes localement compacts (pour un exposé-éclair de moins de 10 pages, voir [**Ka76**] ou [**GGY89**]). A ma connaissance, cette approche n'a été utilisée qu'en traitement de l'image [**GGY89**].

CHAPITRE 2

Une petite théorie statistique de la rareté en Ecologie

Sommaire

Ce travail a été initié dans le cadre du Programme National D'Océanographie Côtière, thème "Séries Longues", et s'est prolongé dans le cadre du Programme National Environnement Côtier (PNEC) qui lui a succédé. Chacune des Séries Longues étudiées a été obtenue en dénombrant les représentants d'un grand nombre (plusieurs centaines) d'espèces marines présentes en un lieu fixe 3 à 4 fois par an, pendant une période de l'ordre de 10 ans (au moins). La longueur de certaines de ces chroniques dépasse maintenant vingt ans.

La difficulté rencontrée dans le traitement exploratoire de ces chroniques réside dans la rareté d'un grand nombre d'espèces. En effet, une méthode comme l'ACP étant très sensible à l'importance des effectifs, ses premiers facteurs ne dépendront que des espèces communes, voire dominantes. L'importance d'autres espèces peu fréquentes, mais d'un grand intérêt écologique (espèces indicatrices de pollution, par exemple) sera donc masquée par l'ACP. Le problème est inverse pour l'Analyse des Correspondances de ce genre de données : les espèces rares peuvent y prendre une importance pathologique. Ce cas a été étudié par Nowak et Bar-Hen [**NB05**]. De leur point de vue, une espèce doit être éliminée si et seulement si elle perturbe excessivement les premières valeurs propres ; leur travail montre que l'exclusion systématique des espèces peu fréquentes n'est pas une bonne stratégie. Une troisième méthode largement utilisée, l'Analyse des Correspondances de la table de présence/absence des espèces, présente aussi de graves défauts. D'abord, l'aspect quantitatif des chroniques (nombre d'individus de chaque espèce) est complètement détruit, ensuite l'existence d'espèces sporadiques vient souvent perturber les résultats qui sont donc naturellement très instables.

La solution serait donc de classer les espèces en deux catégories antinomiques : rares ou communes, puis de les traiter séparément par des méthodes appropriées. Il faut donc pouvoir décider de leur rareté. La caractéristique de ce chapitre sera ainsi l'amalgame de l'exploratoire et du décisionnel.

1. Qu'est-ce que la rareté ?

-Il y a une matière intelligible, qui est en quelque sorte comprimée par sa définition. Orismos *signifie définition ; c'est presque le même mot que* oros *qui veut dire bord. C'est assez remarquable.*

-Définir, c'est dire les frontières, dessiner les frontières ?

-C'est effectivement délimiter les frontières. [**Th93**, p. 110]

La notion de rareté est importante en écologie [**GK97**] : la rareté d'une espèce est en relation avec une multitude d'autres caractéristiques (possibilité d'extinction, évolution, spéciation, diversité génétique, *etc.*). Dans son ouvrage consacré au sujet, Gaston [**Ga94**] met l'accent sur le fait que deux notions distinctes de rareté interviennent dans le langage : une espèce est dite rare soit si elle est représentée par un « petit » effectif d'individus, soit si elle est rarement rencontrée (sens temporel et/ou spatial). De plus, ces notions sont statistiquement liées : une espèce peu abondante est en général rare aussi bien au sens spatial (endémique) qu'au sens temporel [**Ma07, Ga94**].

Gaston [**Ga94**, pp. 158-159] souligne aussi les complications dues à la perturbation des statistiques par les espèces rares, ce qui est proche de la motivation initiale de mon travail : simplifier l'interprétation de l'ACP des séries longues, en éliminant les espèces rares de l'analyse. D'un point de vue pratique, il propose de définir comme rares les espèces appartenant soit au premier quartile de la distribution des abondances (définition **"locale"**, relative à chaque relevé), soit au premier quartile de la distribution du nombre des occurrences (définition **"globale"**, ou spatio-temporelle). Il note cependant (p. 5) que ce genre de définition est peu satisfaisant :

- lorsque la rareté est définie en termes de proportions, une espèce peut changer de catégorie (rare ou commune) sans que ni son effectif ni la structure de la population ne changent notablement ;
- ni les distributions d'abondance, ni les distributions d'occurrence ne présentent généralement de discontinuité permettant de proposer un seuil assimilable à une transition vers la rareté.

Nous sommes, de notre côté, partis d'une définition purement statistique, et de bon sens :

DÉFINITION 1.1. Une espèce est rare si sa probabilité d'échantillonnage est faible.

Pour parvenir à une procédure concrète, il reste à clarifier ce que l'on entend ci-dessus par "probabilité d'échantillonnage", et par "faible". Je le ferai en donnant une suite de définitions emboîtées, jusqu'à parvenir à une notion opérationnelle de rareté locale, qui permettra ensuite de remonter à la rareté globale. La démarche suivie est bien illustrée par cet extrait de *"Do we sleep enough ?"* [**Ben99**] :

> *Does the average man get enough sleep ? What is enough sleep ? What is the average man ? What is "does" ?*

Mais, contrairement à Benchley, je ne dirai rien de Napoléon, ni de ma manière préférée de dormir.

2. Rareté locale et méthodes exploratoires

Il suffit de préciser dans la Définition1.1 ce qu'est une "faible probabilité" pour aboutir à une définition presque utilisable, dépendant du **paramètre de rareté** $\alpha_0 \in [0,1]$.

DÉFINITION 2.1. Une espèce est α_0-rare (resp. α_0-seuil) si sa probabilité de non-prélèvement est supérieure (resp. égale) à la valeur fixée α_0.

Le nombre des espèces rares étant une fonction décroissante du paramètre de rareté, les valeurs pertinentes de α_0 sont plutôt "petites", puisque l'on veut simplifier le dépouillement de l'ACP en les éliminant. Considérons maintenant un relevé comprenant N (aléatoire) organismes, appartenant à l'une des Q (aléatoire) espèces recensées. Soit K_e l'effectif (aléatoire) de l'espèce e, présente sur le site (mais pas forcément dans l'échantillon !) dans une proportion **théorique** π_e.

Nous ferons dans tout ce chapitre l'hypothèse que,
conditionnellement à $N = n$, $K_e \sim \mathcal{B}(n, \pi_e)$ $\qquad (H1)$.

Examinons pour commencer un test "naïf" de l'hypothèse H_e^+ : "l'espèce e est présente" (soit : $\pi_e > 0$), contre sa négation $\overline{H_e^+}$ (soit : $\pi_e = 0$) ; ce test consiste à accepter H_e^+ lorsque $K_e > 0$. Il est on ne peut plus puissant, car son risque de 2^{eme} espèce $P_{\pi_e=0}(K_e > 0)$ est nul, mais son risque de 1^{ere} espèce $P_{\pi_e>0}(K_e = 0) = (1-\pi_e)^n$ dépend de π_e, et sera très important pour une espèce rare ! Ce test n'est donc guère utilisable.

Nous allons néanmoins le ré-introduire immédiatement sous une autre forme, dans le cadre de l'estimation par intervalle. Pour ce, nous allons glisser de l'opposition "présence/absence" (que nous retrouverons dans la Section 3.1) à la notion d'α_0-rareté. Pour une loi binômiale K, définissons la proportion $\pi_0(n)$ telle que

$$\alpha_0 = P_{\pi_0(n)}(K = 0) = (1 - \pi_0(n))^n . \tag{2.1}$$

Nous pouvons sous $(H1)$ reformuler la définition 2.1 d'une manière plus pratique :

Définition 2.2. L'espèce e est α_0-rare (resp. α_0-seuil) si $\pi_e < \pi_0(n)$ (resp. $\pi_e = \pi_0(n)$).

Mais l'assertion $\pi_e < \pi_0(n)$ n'équivaut pas à l'événement $\widehat{\pi_e} < \pi_0(n)$, où $\widehat{\pi_e}$ désigne l'estimateur du maximum de vraisemblance de π_e. Une manière de parvenir à une décision est d'attribuer une **crédibilité** à l'événement $\pi_e < \pi_0(n)$, au vu d'une réalisation k_e de K_e. Pour ce, nous avons proposé **[P4]** de comparer $\pi_0(n)$ à la **limite supérieure de confiance [UB73]** de π_e, à un niveau donné η_0. C'est une statistique $\widehat{\pi_e}^*(k_e, n)$ vérifiant :

$$\underset{\pi\in[0,1]}{Inf}\ P_\pi(\widehat{\pi_e}^* \geq \pi) = \eta_0.$$

Par exemple, si l'on choisit le niveau 95%, il y a une probabilité supérieure à $\eta_0 = 0.95$ pour que $\widehat{\pi_e}^*(k_e, n)$ dépasse la "vraie" proportion π_e. Si $\widehat{\pi_e}^*(k_e, n) \leq \pi_0(n)$, il est donc très probable que e soit α_0-rare. Nous pouvons maintenant poser une définition pleinement opérationnelle :

Définition 2.3. L'espèce e est (α_0, η_0)-rare si $\widehat{\pi_e}^*(k_e, n) \leq \pi_0(n)$.

Rappelons que dans ce couple, α_0 doit être "petit" et η_0 "grand". Après avoir fixé (arbitrairement pour l'instant) ces deux paramètres, on peut procéder à la **sélection locale** des espèces, en éliminant de chaque relevé celles qui sont (α_0, η_0)-rares.

Remarque 2.4. Le nombre d'espèces (α_0, η_0)-rares dans un relevé est une fonction décroissante des deux paramètres **[P12*]**.

Cette méthode a d'abord été appliquée à l'ACP de relevés benthiques effectués sur le site de Pierre Noire, en Baie de Morlaix **[P4]**. Les données consistaient en 112 relevés s'étalant d'avril 1977 à Décembre 1991 ; durant cette période, les biologistes ont recensé 421 espèces dans les sédiments prélevés. Nous avons trouvé que la majorité (70.5%) de ces espèces étaient rares dans **tous** les relevés, donc sans influence concrète sur l'ACP. On peut constater dans cet article que les interdistances de Bhattacharya **[P4]** entre relevés bruts d'une part, et celles entre relevés restreints aux espèces communes d'autre part, étaient quasiment identiques. Il est donc légitime de restreindre l'analyse aux espèces communes, ce qui en simplifie considérablement le dépouillement, sans perte d'information notable.

2.1. L'indice de présence/absence lissé.

Les calculs développés dans **[P4]** nous ont permis d'introduire dans **[P6]** un indice de présence/absence lissé (***IPAL***) γ_{α_0} prenant ses valeurs dans $[\alpha_0, 1]$. Il est défini à partir de la fonction :

$$\begin{aligned} \eta_{\alpha_0}(k_e, n) : &= \sum_{i=1+k_e}^{n} \binom{n}{i} \pi_0(n)^i (1 - \pi_0(n))^{n-i} \\ &= \beta(k_e + 1, n - k_e)(\pi_0(n)) \end{aligned} \tag{2.2}$$

où $\beta(i, j)(x) := \int_0^x t^{i-1}(1-t)^{j-1}\,dt$ désigne une fonction Bêta- incomplète. Si l'on se rappelle que la limite supérieure de confiance de niveau η_0 pour π_e est l'unique solution en p de l'équation $\eta_0 = \beta(k_e + 1, n - k_e)(p)$, on peut encore écrire **[P12*]** :

Définition 2.5. L'espèce e est (α_0, η_0)-rare si $\eta_{\alpha_0}(k_e, n) > \eta_0$.

Grâce à la formule (2.2), notre problème de décision se ramène finalement au calcul d'une valeur d'une fonction Bêta-incomplète, alors que la Définition 2.3 nous obligeait à inverser cette fonction.

REMARQUE 2.6. On peut montrer [**P12*, P13**] que, si $\eta_0 \geq 1-\alpha_0$, il n'y a pas de sélection possible. En effet, dans ce cas

$$\eta_{\alpha_0}(0,n) = 1-\alpha_0$$
$$\eta_{\alpha_0}(k,n) < 1-\alpha_0,\ \forall k>0$$

d'après la formule 2.2. Donc, dès que l'espèce e est présente dans un relevé, $\eta_{\alpha_0}(k_e,n) < \eta_0$. Selon la définition 2.5, aucune espèce présente ne peut alors être (α_0,η_0)-rare, dans aucun relevé.

L'indice *IPAL* introduit dans [**P6**] est $\gamma_{\alpha_0}(k,n) := 1-\eta_{\alpha_0}(k,n)$. Si α_0 est "petit", il peut être interprété comme un indice de présence/absence lissé, car $\gamma_{\alpha_0}(0,n) = \alpha_0$ et $\gamma_{\alpha_0}(n,n) = 1$. Mais n est en général grand, et il serait souhaitable qu'une espèce soit considérée comme présente (valeur de γ_{α_0} proche de un) à partir d'un effectif raisonnable(quelques représentants) ! Que se passe-t-il donc pour des valeurs intermédiaires de k ? En posant $\lambda_0 := -\ln(\alpha_0) \in \mathbb{R}^+$, on peut démontrer [**P12***] la convergence de la loi binômiale vers la loi de Poisson :

$$\gamma_{\alpha_0}(k,n) = P(\mathcal{B}(n,\pi_0(n)) \leq k) \underset{n\to\infty}{\longrightarrow} P(\mathcal{P}(\lambda_0) \leq k) := \widehat{\gamma_{\alpha_0}}(k).$$

Comme on peut le voir sur la figure 2.1, cette convergence est rapide lorsque α_0 n'est pas trop petit. Cette figure montre que, pour une valeur courante de l'effectif n (pour les relevés de la baie de Morlaix, $n \approx 3750$), on peut très bien substituer $\hat{\gamma}$ à γ. Par conséquent, dans la pratique, la courbe $\gamma_{\alpha_0}(\bullet,n)$ ne dépend pas de n, et les espèces seront considérées comme présentes pour des effectif modestes, car $\widehat{\gamma_{\alpha_0}}(k) \approx 1$ pour de petites valeurs de k (de l'ordre de la vingtaine).

REMARQUE. L'indice γ_{α_0} est une **distribution de possibilité [Du06]**, ce qui ouvre des horizons que nous n'avons malheureusement pas explorés.

Nous avons montré dans [**P6**] que *IPAL* se substitue avantageusement à l'indice de présence/absence pour l'Analyse des Correspondances. En effet, contrairement à la variante classique en 0/1, il recode les effectifs en conservant partiellement la nature quantitative des données (voir figure 2.1). Cependant, dans l'AFC en *IPAL*, ni les espèces toujours dominantes (pour lesquelles γ_{α_0}vaut toujours 1), ni les espèces toujours rares (pour lesquelles γ_{α_0}vaut toujours α_0) ne jouent de rôle. Le choix du couple (α_0,η_0) est donc crucial.

2.1.1. *Loi asymptotique de IPAL pour une espèce seuil.*

Soit $\theta_i := F_{\mathcal{P}(\lambda_0)}(i) = P(\mathcal{P}(\lambda_0) \leq i)$, $\Theta_{\alpha_0} := \{\theta_0,\cdots,\theta_i,\cdots\}$, et $\Re_{\alpha_0}$ la loi de probabilité supportée par Θ_{α_0}, de densité :

$$P(\Re_{\alpha_0} = \theta_k) = \frac{e^{-\lambda_0}\lambda_0^k}{k!}.$$

Remarquons que la forme de $\Re_{\alpha_0}$ aussi bien que son support, dépendent étroitement de α_0 (voir la figure 2.2). Par construction, $\Re_{\alpha_0} = F_{\mathcal{P}(\lambda_0)}(\mathcal{P}(\lambda_0))$ est la "loi uniforme" sur Θ_{α_0}(voir la figure 2.3). On démontre enfin la proposition suivante [**P12***] :

PROPOSITION 2.7. *Si l'espèce e est* α_0-seuil, *la variable aléatoire* $\gamma_{\alpha_0}(K_e,n)$ *converge en loi vers* $\Re_{\alpha_0}$ *lorsque n tend vers l'infini.*

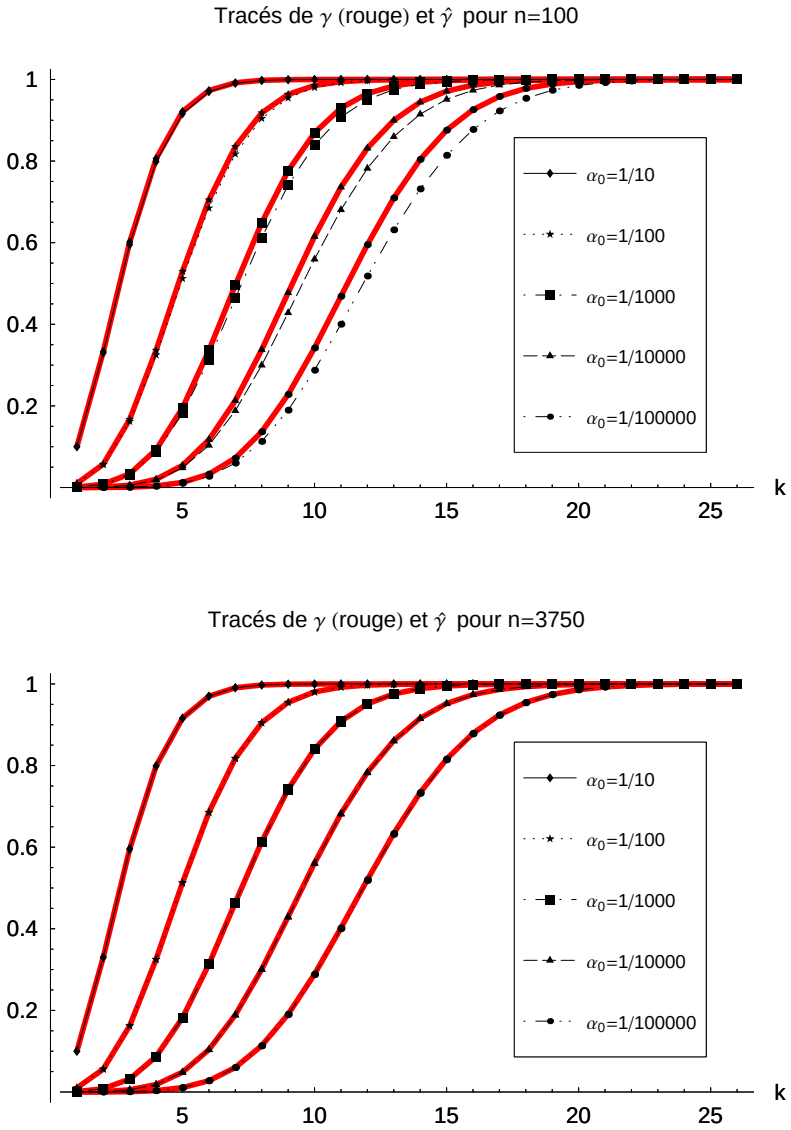
Cette proposition va nous permettre de nous attaquer à la rareté globale.

3. De la rareté locale à la rareté globale

Le moment est venu de revenir aux deux notions emboîtées de rareté évoquées par Gaston [**Ga94**] : une espèce est rare soit si elle est représentée par un « petit » effectif d'individus, soit si elle est rarement rencontrée. La première notion correspond à la α_0-rareté locale, évaluée ci-dessus via la (α_0,η_0)-rareté. La deuxième notion va être associée à la statistique, espèce par espèce, de *IPAL*.

DÉFINITION 3.1. L'espèce e est globalement α_0-rare si elle est généralement plus rare qu'une espèce α_0-seuil.

FIG. 2.1. ***Les fonctions*** $\gamma_{\alpha_0}(k,n)$ ***et*** $\widehat{\gamma_{\alpha_0}}(k)$***, pour différentes valeurs de*** n ***et*** α_0***.***



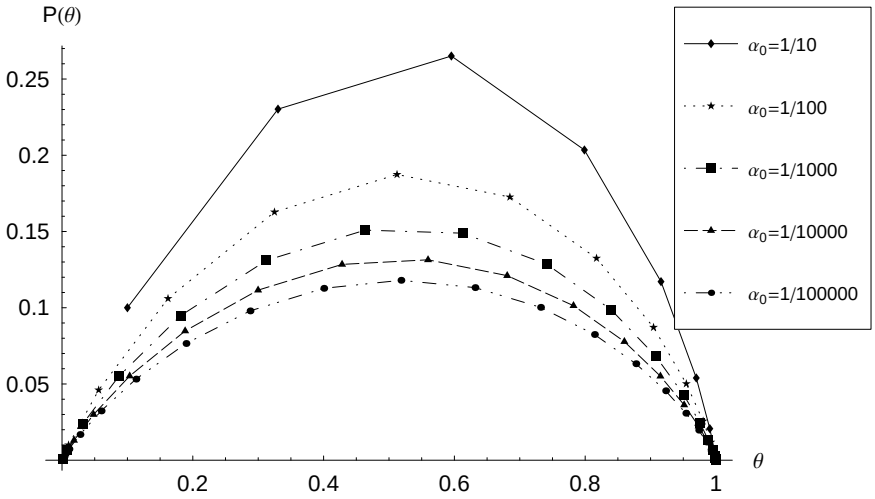
Le terme "généralement" a ici un sens purement statistique. A partir des effectifs $\left\{k_e^1, \cdots k_e^d, \cdots k_e^D\right\}$ de e dans D relevés, nous pouvons construire la distribution empirique $\Gamma_{\alpha_0}^e$ de $\gamma_{\alpha_0}(K_e, N)$ et la comparer à celle de $\Re_{\alpha_0}$, afin de voir si e est plus rare que cette espèce théorique. Remarquons que cela exige que l'effectif de toute espèce rare ou seuil soit stationnaire (cas temporel) ou homogène isotrope (cas spatial).

Nous accepterons dans la suite de cette section une hypothèse plus forte : pour toute espèce **e** *α_0-rare ou α_0-seuil, $\left\{k_e^1, \cdots k_e^d, \cdots k_e^D\right\}$ forme un D-échantillon i.i.d. d'une loi dont les deux premiers moments sont finis (**H2**).*

Cette hypothèse nous évite les ennuis liés à la corrélation éventuelle des observations. Il est certes possible d'estimer correctement la fonction de répartition de données corrélées [**EV06, Bo99**], mais le problème paraît beaucoup plus délicat pour les tests d'adéquation correspondants... L'hypothèse H2 de faible interaction spatio-temporelle semble acceptable dans le cas d'espèces rares, alors qu'elle serait intenable pour les espèces communes, dont le comportement est souvent densité-dépendant [**P14**].

Dans [**P12***] sont proposés deux tests de rareté globale, fondés sur la comparaison entre $\Gamma_{\alpha_0}^e$ et $\Re_{\alpha_0}$. Le premier est basé sur le théorème de la limite centrale (TLC). Il se justifie donc bien

FIG. 2.2. ***La densité de $\Re_{\alpha_0}$ pour différentes valeurs du paramètre de rareté.***



pour des relevés d'effectif important, car les moments de la loi asymptotique de *IPAL* sous **H2** sont substitués à ses vrais moments.

DÉFINITION 3.2. L'espèce e est globalement α_0-rare au niveau de confiance ζ et au sens du TLC si :

$$\frac{\sum_{i=1}^{D}\left(\gamma_{\alpha_0}\left(k_e^i, n_i\right) - E\left(\Re_{\alpha_0}\right)\right)}{\sqrt{D\,Var\left(\Re_{\alpha_0}\right)}} < \Phi(\zeta)$$

où Φ désigne la fonction de répartition (**f.r.**) de la loi normale centrée réduite.

Le deuxième test est basé sur la dominance stochastique éventuelle de la f.r. $\Gamma^e_{\alpha_0}$ par celle de $\Re_{\alpha_0}$; il fonctionne donc bien quand le nombre de relevés est important. Il est bon de rappeler la définition de cette relation entre distributions.

DÉFINITION 3.3. **[Le55]** La variable aléatoire X (de f.r. F_X) domine stochastiquement Y (de f.r. F_Y) si, pour tout t, $F_X(t) \leq F_Y(t)$. Cet ordre partiel est (contre-intuitivement) noté $X \ll Y$.

Nous pouvons maintenant donner une deuxième définition de la rareté globale, associée à la notion d'ordre stochastique.

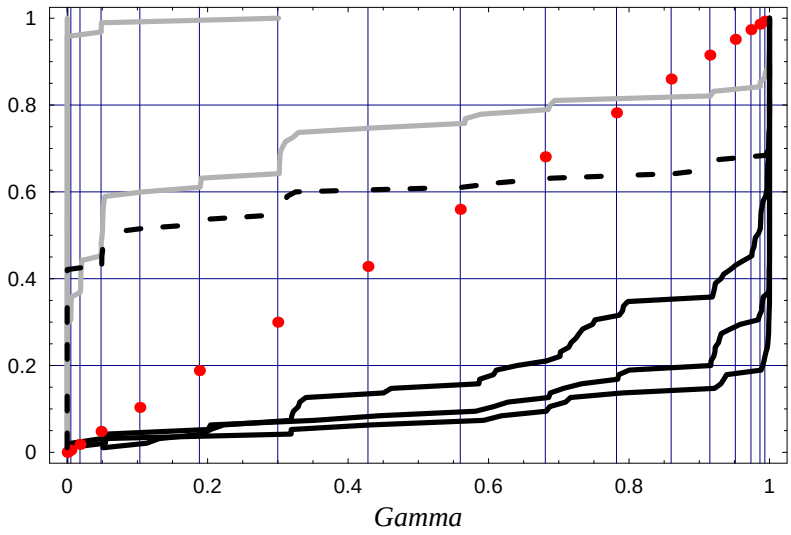
DÉFINITION 3.4. L'espèce e est globalement α_0-rare (resp. seuil) au sens de l'ordre stochastique si $\Gamma^e_{\alpha_0} \underset{\neq}{\gg} \Re_{\alpha_0}$ (resp. $\Gamma^e_{\alpha_0} = \Re_{\alpha_0}$).

Plutôt que d'exposer longuement la mise en oeuvre des tests associés, détaillée dans [**P12***], nous illustrerons cette méthode par quelques résultats concrets, obtenus pour $\alpha_0 = 0.0001$. Sur la figure 2.3 sont représentées les f.r. $\{\Gamma^{e_1}_{0.0001}, \cdots, \Gamma^{e_7}_{0.0001}\}$ de sept espèces benthiques collectées sur un autre site, celui de la Baie de Seine (95 relevés). Les deux tests indiquaient que les espèces associées aux courbes grises étaient rares et celle dont la f.r. est en tirets était 0.0001-seuil. Les autres étaient communes. Nous avons trouvé qu'environ 90% des 115 espèces rencontrées sur ce site très pollué étaient globalement rares. Pour la même valeur du paramètre de rareté, environ 95% des 421 espèces de Pierre Noire l'étaient.

3.1. Choix du paramètre de rareté.

La valeur arbitraire $\alpha_0 = 0.0001$ utilisée ci-dessus était-elle vraiment pertinente ? Et sinon, est-il possible d'optimiser ce paramètre ? La réponse n'est pas la même selon que l'on s'intéresse à la rareté globale, dans une optique décisionnelle, ou à la rareté locale, dans une optique purement exploratoire. Le deuxième point de vue sera traité dans la section suivante. Quand au premier,

FIG. 2.3. ***Fonctions de répartition de IPAL pour des espèces globalement rares (gris) ou communes (noir). La f.r. tiretée correspond à une espèce seuil, et les points rouges à la f.r. de*** $\Re_{0.0001}$***. Les barres verticales correspondent à*** $\{\theta_0, \cdots, \theta_{16}\}$***.***



nous l'avons seulement abordé dans une communication non publiée (C. Manté, J.C. Dauvin and J.P. Durbec, *Rarity and non-sampling*, Mathematical methods in Oceanography - deterministic and stochastic aspects, Marseille, 13-17 december 1999). Afin de compléter l'exposé, nous allons poursuivre ici ce travail inachevé.

Remarquons d'abord que le problème global est naturellement binaire, puisqu'il s'agit de décider si une espèce peut être considérée comme rare (ou non) au vu de sa **possibilité** de présence (d'absence) dans une suite de relevés. Supposons que l'espèce e se trouve dans un relevé, dans la proportion $\widehat{\pi_e} := \frac{k_e}{n}$. La formule 2.1 nous permet d'associer à cette proportion la possibilité $\alpha_e := (1 - \widehat{\pi_e})^n$ de l'événement $K_e = 0$, c'est-à-dire du non-prélèvement de cette espèce. Nous voudrions choisir le paramètre de rareté de façon à ce que l'événement $\alpha_e \geq \alpha_0$ soit un bon prédicteur de $K_e = 0$. Mais, vu que $k_e = 0 \Leftrightarrow \widehat{\pi_e} = 0$, si l'on ne dispose que d'un échantillon, on conclura que l'effectif est un excellent prédicteur de lui-même... Pour contourner la circularité de ce raisonnement, il faut pouvoir estimer **indépendamment** π_e et k_e.

Cela est possible si nous disposons pour chaque relevé de **réplicats** (c'est-à-dire d'autres échantillons prélevés dans les mêmes conditions), ce qui fut le cas à Pierre Noire [**P4**]. Chaque relevé était en effet obtenu en cumulant les organismes dénombrés séparément dans dix bennes distinctes ; nous avons considéré ces bennes comme autant de réplicats. Un relevé correspond donc ici à un tableau T d'effectifs, à dix colonnes et E (aléatoire) lignes ; aucune ligne n'est vide. A la case (e, r) de ce tableau, nous associons trois statistiques :

$$n^r := \sum_{j \in \{1,\cdots,E\}} T_j^r$$

$$\widetilde{\pi_e^r} := \frac{\sum_{i \in \{1,\cdots,10\}-\{r\}} T_e^i}{\sum_{j \in \{1,\cdots,E\}} \sum_{i \in \{1,\cdots,10\}-\{r\}} T_j^i}$$

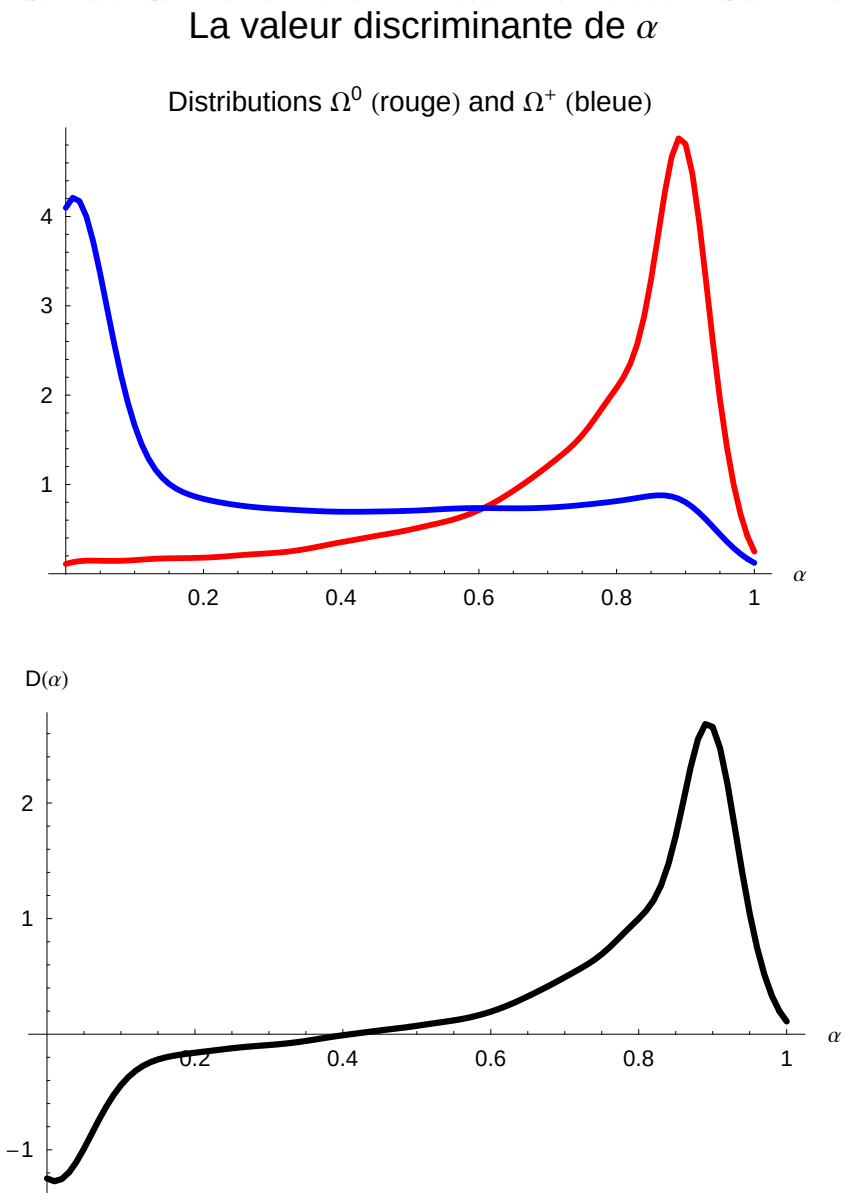
$$\alpha_e^r := \left(1 - \widetilde{\pi_e^r}\right)^{n^r} .$$

La v.a. n^r est l'effectif du réplicat r, et $\widetilde{\pi_e^r}$ est l'estimation de π_e obtenue à partir des autres réplicats ; lorsque $\alpha_e^r \approx 1$, nous devrions avoir $k_e = T_e^r \approx 0$. Mais ces deux variables sont maintenant indépendamment échantillonnées ; par conséquent, nous pouvons décomposer l'ensemble des valeurs de α_e^r en deux parties complémentaires : les "absences" $\Omega^0 := \{\alpha_e^r : T_e^r = 0\}$ et les "présences" $\Omega^+ := \{\alpha_e^r : T_e^r > 0\}$. Prévoir l'état (présente/absente) de chaque espèce dans chacune des cases de chaque tableau à partir de α_e^r est un problème classique d'analyse discriminante

décisionnelle [**DM93**]. Si nous choisissons une fonction de coût symétrique, la solution est racine de l'équation :

$$D(\alpha) := \nu\left(\Omega^0\right) f_{\Omega^0}(\alpha) - \nu\left(\Omega^+\right) f_{\Omega^+}(\alpha) = 0$$

FIG. 2.4. ***En haut : estimation des densités*** f_{Ω^0} ***et*** f_{Ω^+} ***dans le cas des données de Pierre Noire. En bas : la fonction discriminante*** $D(\alpha)$**.**



où ν désigne la mesure de dénombrement, et $f_\bullet$ la densité de probabilité de l'indice dans chaque ensemble.

Le résultat obtenu avec les données de Pierre Noire est représenté sur la figure 2.4. Les densités sont ici estimées grâce au package *Mathematica* de B. Gress (`http://student.ucr.edu/~gressb01/Nonparametrix_m.html`). Nous avons utilisé le noyau d'Epanechnikov, la largeur de bande étant déterminée par la règle de Silverman [**Si86**, p. 48].

On notera que, au sein d'un même relevé, les valeurs $\left\{\alpha_e^1, \cdots, \alpha_e^{10}\right\}$ ne sont pas indépendantes. Grâce à [**EV06**], nous savons cependant que ce n'est pas un handicap pour l'estimateur à noyau de la densité. La fonction discriminante est donc elle aussi correctement estimée. On peut voir clairement sur la figure 2.4 qu'elle s'annule pour $\alpha_{Disc} \approx 0.4$. Dans le cas des "absences", le pourcentage de bonnes prédictions (obtenues avec la règle $\alpha \geq \alpha_{Disc}$) est de 92%, mais il est

seulement de 65% pour les "présences". Environ 35% des "présences" sont donc mal prévues, mais 80% de ces erreurs correspondent ici à des effectifs de un, à des présences très improbables.

A Pierre Noire, la même procédure donne un pourcentage d'espèces globalement α_{Disc}-rares proche de 80% [**P14**], ce qui paraît plus raisonnable que les 95% rencontrés dans la section précédente, pour la valeur arbitraire $\alpha_0 = 0.0001$.

4. Sélection automatique des espèces pour l'ACP

Dans [**P13**], nous avons proposé une méthode différente d'optimisation du paramètre de rareté. L'objectif est cette fois purement exploratoire : il s'agit de trouver un couple (α_0, η_0) tel que **l'essentiel de la structure du nuage** des relevés soit fidèlement restitué par l'analyse des espèces communes, tout en sélectionnant localement le moins d'espèces possible. On entend ici par « l'essentiel de la structure du nuage », la projection de celui-ci sur l'espace de ses k (fixé) premières composantes principales.

Nous avons ici mis à profit le fait que η_0 peut être choisi en fonction de α_0 et d'un niveau de confiance τ fixé (voir ci-dessous). On peut ainsi optimiser la procédure par rapport au seul paramètre de rareté, au lieu du couple (α_0, η_0), ce qui accélère considérablement l'algorithme de recherche de l'optimum.

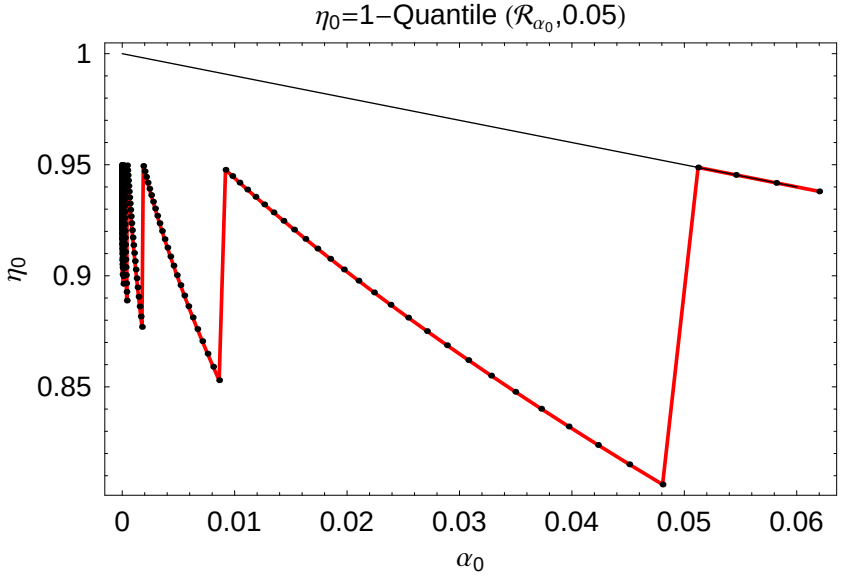
4.1. Construction de la courbe $\eta(\alpha_0)$.

Nous avons vu qu'à une valeur donnée du paramètre de rareté correspond la distribution théorique $\Re_{\alpha_0}$ des espèces-seuils, qui est d'après la Proposition 2.7 la loi asymptotique de $\gamma_{\alpha_0}(K_e, n)$ pour de telles espèces. Or, la Définition 2.5 nous dit que e est (α_0, η_0)-rare si $\gamma_{\alpha_0}(k_e, n) < 1 - \eta_0$, cette dernière quantité devant être "petite". Par conséquent, il est naturel de prendre pour valeur de η_0 une "grande " valeur de la variable aléatoire $1 - \gamma_{\alpha_0}$. Cela nous a conduits à proposer [**P12***] le choix :

$$\eta_0 = \eta(\alpha_0) := 1 - Quantile\,[\Re_{\alpha_0}, \tau]$$

avec τ fixé et "petit" - classiquement, 0.05.

FIG. 2.5. ***La valeur optimale de*** η_0 ***en fonction de*** α_0 ***et de*** $\tau = 0.05$ ***, échantillonnée en 500 point. En noir : la droite d'équation*** $y = 1 - x$.



Il est utile de remarquer que $\eta(\alpha_0)$ n'est pas une fonction monotone de α_0. Comme on peut le voir sur la figure 2.5, cette fonction n'est pas continue non plus, et la courbe associée ne possède pas de tangente en zéro. Cette figure illustre également le fait que, lorsque $\alpha_0 \geq \tau$, $\eta(\alpha_0) = 1 - \alpha_0$ [**P12*, P13**]. Or, nous avons vu (remarque 2.6) qu'aucune sélection n'est alors possible. Ce cas de figure est donc à écarter d'emblée, et l'on n'explorera que la courbe $(\alpha, \eta(\alpha)) : \alpha \in]0, \tau[$.

4.2. Le critère de sélection.

Pour une valeur donnée de α_0, le critère de fidélité choisi est la distance de Krzanowski [**Kr87**] entre d'une part le nuage projeté des relevés complets (correspondant à $\alpha = 0$: conservation de toutes les espèces), d'autre part le nuage projeté des relevés réduits aux espèces localement communes, avec $\alpha = \alpha_0$.

La distance de Krzanowski est une distance procustéenne entre deux nuages de points de même effectif (ici : les relevés) ; elle est invariante par rotation, translation et réflexion, mais n'est pas une fonction croissante de α. Elle est seulement croissante "en tendance", dans la mesure où elle est nulle pour $\alpha = 0$, et maximale pour $\alpha = 1$ (tableau nul). La raison en est que le nombre d'espèces éliminées est une fonction décroissante de α_0 et η_0 (remarque 2.4), alors que comme le montre la figure 2.5, $\eta(\alpha_0)$ est seulement une fonction décroissante par morceaux du paramètre de rareté.

4.3. Optimisation du critère.

Notre problème est complètement trivial si nous cherchons un optimum vis-à-vis de la seule distance de Krzanowski : il suffirait de choisir $\alpha_0 = 0$, donc d'analyser le tableau brut. Mais il s'agit pour nous de "simplifier" l'ACP. Ici, cela signifiera **minimiser son coût de calcul** (ou complexité), qui servira de pénalisation. Nous assimilons ce coût à celui de la décomposition en valeurs singulières associée, donné par Golub & van Loan [**GL96**] en fonction des dimensions du tableau.

Pratiquement, nous nous donnons une suite $0 < \alpha^1 < \cdots \alpha^p < \tau$ de valeurs du paramètre de rareté ; pour chaque valeur α^i, nous éliminons les espèces $\left(\alpha^i, \eta\left(\alpha^i\right)\right)$-rares du tableau complet T^0. Cela nous donne un tableau "simplifié" T^i , dans lequel une colonne (correspondant à une espèce) est détruite dès lors qu'elle ne contient que des zéros. Le coût de calcul $0 < C\left(T^i\right) \leq C\left(T^0\right)$ est décroissant en tendance, vu que le nombre de colonnes du tableau à décomposer décroît lui-même en tendance. Nous traçons ensuite le diagramme :

$$\left\{ \left(D_{Kr}\left(T^0, T^i\right), \frac{C\left(T^i\right)}{C\left(T^0\right)} \right), i \in \{1, \cdots, p\} \right\}.$$

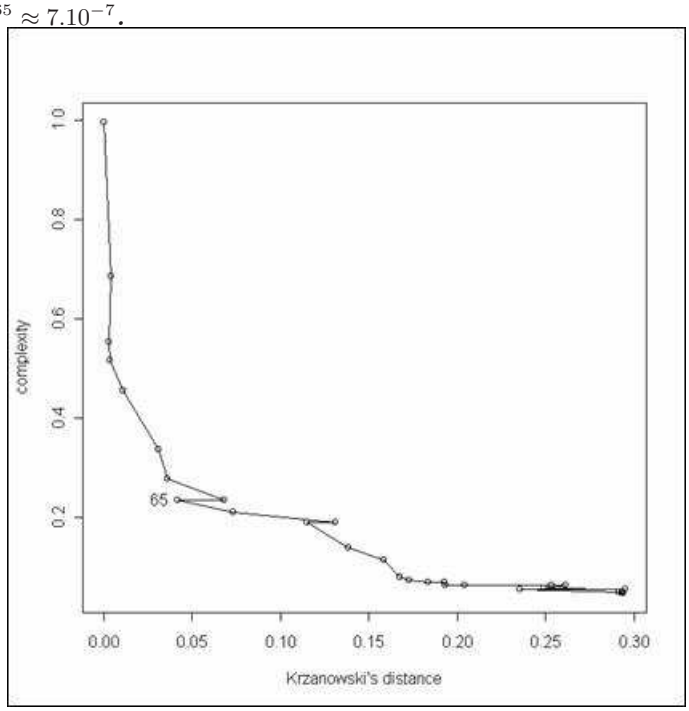
Une "bonne" valeur α_{Opt} du paramètre devrait correspondre à un "coude" sur ce diagramme : pour $\alpha^i < \alpha_{Opt}$ le coût de calcul augmente rapidement, alors que la distance $D_{Kr}\left(T^0, T^i\right)$ est petite et ne peut guère diminuer, la réciproque se produisant pour $\alpha^i > \alpha_{Opt}$: le coût est faible et diminue lentement, alors que la distance augmente rapidement.

Cette méthode a été introduite dans le stage de DEA de J. Claudet (2002), pour traiter des relevés benthiques provenant du port d'Alger. C'étaient des données spatio-temporelles : 30 stations couvrant les trois bassins du port ont été échantillonnées quatre fois (au plus) dans l'année 1983-1984. Deux-cent soixante et une espèces ont été rencontrées au cours de ces campagnes, et le tableau complet était du type 97×261. Le diagramme correspondant à la projection dans l'espace des $k = 5$ premières composantes est représenté sur la figure 2.6 extraite de [**P13**], que le lecteur pourra consulter pour plus de détails. Seules 71 espèces sont conservées, avec $\eta_{Opt} = \eta(\alpha_{Opt}) \approx 0.94$.

5. En guise de conclusion

Les méthodes proposées ci-dessus nous permettent de diviser l'ensemble des espèces collectées en deux (ou trois) catégories exclusives : communes ou rares (et éventuellement seuils). Cette possibilité devrait être utile aux écologistes, pour qui la rareté n'est pas seulement intéressante en soi [**Ga94, GK97**], mais est un facteur à prendre en compte dans les modèles de distribution d'abondance dans les communautés. On observe en effet que si les effectifs des espèces communes sont généralement bien ajustés par une distribution log-normale, les effectifs des espèces rares sont plutôt distribués selon une loi de puissance, ou suivant une distribution "log series" de Fisher [**MH03, UO04**].

FIG. 2.6. ***Le diagramme fidélité × coût des données du port d'Alger, pour k=5 composantes. Les points sont liés en fonction du paramètre de rareté. L'optimum choisi est :*** $\alpha^{65} \approx 7.10^{-7}$.



Que l'on procède localement (optique exploratoire) ou globalement (optique décisionnelle), tout l'échafaudage repose ici sur la notion d'α_0-rareté. Il est donc capital d'utiliser une valeur pertinente du paramètre de rareté. Nous avons vu que dans le cas de l'ACP (voir Section 4), il faut prendre de petites valeurs de α_0, alors que dans le cas décisionnel (rareté globale : voir Section 3.1) on est conduit à choisir des valeurs plus importantes. Le cas de l'AFC en ***IPAL*** est au carrefour des méthodes précédentes : c'est une méthode exploratoire pour traiter des présences/absences lissées. Nous avons montré **[P6]** que cette méthode peut donner de meilleurs résultats que l'AFC du tableau en 0/1 correspondant, mais comment optimiser α_0 ? Krzanowski **[Kr93]** a proposé une variante de son critère adaptée à l'AFC, mais quel tableau de référence devrions-nous choisir ? Le tableau en 0/1 ayant une structure très instable due aux espèces rares, il ne semble pas être un bon candidat. Il serait sans doute mieux, lorsque cela est possible (existence de réplicats), de prendre pour référence le tableau obtenu avec le paramètre de rareté α_{Disc}, en raison de sa "robustesse". On procédera alors comme ci-dessus, avec $\alpha \in]0, \min(\tau, \alpha_{Disc})[$, ce qui permettra d'avoir une analyse aussi fine que possible, sans trop s'éloigner du tableau initial T^{Disc}. Notons qu'ici, le tableau de référence correspond à un coût minimal, car on peut en éliminer toute espèce pour laquelle les valeurs de l'indice sont identiques (colonne constante). Lorsque $\alpha \to 0$, l'indice est de plus en plus sensible aux différences d'effectifs (voir figure 2.1) ; par conséquent, le nombre de colonnes non constantes augmente, et avec lui la complexité de la SVD du tableau.

CHAPITRE 3

Analyse en Composantes Principales de mesures absolument continues : applications en Sédimentologie et en Ecologie

Sommaire

Il est très fréquent dans les Sciences de l'Environnement que des données issues d'appareils de mesure se présentent spontanément sous la forme de courbes cumulées assimilables à des fonctions de répartition. On citera par exemple les courbes de démagnétisation (paléomagnétisme), les courbes de porosité (structure des roches ou des sols) et les courbes granulométriques (structure des sédiments et des granulats, médecine), dont nous allons parler. La principale originalité de notre méthode est de reposer sur l'existence d'une probabilité de référence permettant d'orienter l'analyse, et de tenir compte d'un changement éventuel de l'échelle de mesure. Nous illustrerons notre propos par deux applications, l'une en sédimentologie, l'autre en écologie benthique.

1. Une application en sédimentologie

Les données granulométriques se présentent sous la forme de courbes cumulées, dont l'abscisse correspond à la taille des particules composant le sédiment. Elles sont couramment utilisées par les géographes et les géologues pour caractériser les sédiments en fonction de la taille des grains les constituant (argile, sable, gravier, *etc.*) ; leur allure générale est d'autre part étroitement liée aux conditions de dépôt dans le milieu, par exemple aux modalités du transport sédimentaire.

1.1. Caractéristiques des courbes.

Quel que soit l'appareillage utilisé (tamis, pipette, compteur Coulter, Sédigraphe, compteur laser, *etc.*) elles sont discrétisées, les tailles mesurées étant imposées par le constructeur. Bien que leur graphe présente toutes les caractéristiques d'une fonction de répartition (**f.r.**) empirique, elles n'en sont pas : on n'a accès ni à la taille de chaque grain, ni même au nombre de grains appartenant à une même classe de tailles. Les tailles échantillonnées croissent généralement suivant une progression géométrique, car ces courbes sont classiquement représentées suivant une échelle logarithmique. Cette échelle, dite échelle ϕ, s'impose aussi bien pour des raisons pratiques (les petites particules sont beaucoup plus nombreuses que les grosses) que théoriques (Kolmogorov a démontré en 1940 que la taille de particules produites par concassage suit asymptotiquement une loi log-normale).

1.2. Approximation de la densité pour la sédimentologie [Kru34, P8*].

Les trois caractéristiques des courbes granulométriques (structure de f.r., discrétisation et existence d'une échelle) ont été prises en compte par Krumbein dans son célèbre article de 1934 [**Kru34**]. Il y fait remarquer que l'histogramme associé à la discrétisation imposée par un système de tamis est un estimateur très instable de la densité, car il dépend grandement des tailles échantillonnées. Il propose à la place de l'approcher en dérivant graphiquement la f.r. (après avoir judicieusement interpolé cette fonction) ; de plus, sa méthode tient compte de la

nature logarithmique de l'échelle ϕ. Notons que le premier estimateur non paramétrique de la densité (autre que l'histogramme) fut introduit en Statistique par Fix et Hodges en 1951 [**Si86**, p. 2] ! Dans ce chapitre, nous allons approfondir le point de vue de Krumbein, dans la perspective d'analyser des données granulométriques par une méthode exploratoire adaptée.

Nous nous sommes d'abord attaqués [**P8***] au problème de la dérivation, en exploitant le fait qu'une courbe granulométrique F est **physiquement** attachée à une **mesure** positive ν : si x et y $(x < y)$ sont deux tailles échantillonnées, $F(y) - F(x) = \nu(]x, y])$ correspond réellement à la masse (ou au nombre) des particules dont la taille appartient à $]x, y]$. La distribution des tailles peut donc être caractérisée par la dérivée usuelle $\frac{dF}{dx}$ (si elle existe), mais remarquons que cette fonction est modifiée lorsque l'on change d'échelle (voir la figure 1 de [**P15***] ou la figure 3.1 ci-après). La ressemblance entre deux sédiments dépendrait donc de l'échelle utilisée, ce qui n'est pas satisfaisant ! Krumbein évoque déjà ce problème dans l'article cité :

> *... the diameters are usually plotted on a logarithmic scale, and here again the choice is immaterial, providing due cognizance is taken of the change in the shape of the curve induced by transformation to a logarithmic scale.* [**Kru34**, p. 66]

Mais ne perdons pas de vue que $\frac{dF}{dx}$ est la limite quand $\varepsilon \to 0$ de $\frac{\nu(]x,x+\varepsilon])}{\lambda(]x,x+\varepsilon])}$, où λ désigne la mesure de Lebesgue : c'est la **dérivée de Radon-Nikodym** de ν relativement à la **mesure de référence** λ, définie dès que ν est dominée par λ. Étant donné que nous avons affaire à des mesures, il est naturel de faire appel à cette notion classique en théorie de l'intégration. Cette formulation nous évite de privilégier la mesure de Lebesgue, ce qui se justifie largement, comme nous le verrons dans la Section1.3.

Dans [**P8***], nous avons traité le problème général de l'approximation hilbertienne de densités de Radon-Nikodym. Nous nous sommes limités au cas où les mesures (éventuellement signées) sont à support dans un intervalle borné $[a, b]$, et où la mesure de référence est une probabilité μ équivalente à λ sur cet intervalle. On suppose de plus que :

$$f := \frac{d\nu}{d\mu} \in \mathbf{L}^2_\mu := \mathbf{L}^2_\mu([a, b]).$$

L'espace usuel des fonctions de carré intégrable sur $[a, b]$ sera simplement noté $\mathbf{L}^2$.

Considérons l'opérateur de $\mathcal{L}\left(\mathbf{L}^2_\mu, \mathbf{L}^2\right)$ défini par : $\widetilde{\Im}(f) = F$. Nous avons montré qu'il ne possède pas d'inverse généralisé borné mais que, si l'on impose à F d'appartenir à un sous-espace bien choisi de $\mathbf{L}^2$, noté H_μ, le problème de l'inversion de l'opérateur restreint $\Im \in \mathcal{L}\left(\mathbf{L}^2_\mu, H_\mu\right)$ est bien posé. En d'autres termes, on peut toujours trouver f de carré μ-intégrable vérifiant $\Im(f) = F \in H_\mu$. L'espace H_μ est un espace de Hilbert à noyau reproduisant, dont le noyau K^μ dépend de a, b, et μ. Par construction, $\Im$ est unitaire, et $\Im^{-1} = \Im^*$ est l'opérateur de dérivation de Radon-Nikodym relatif à μ.

Il faut bien vite sortir de ce cadre abstrait, car les valeurs de F ne sont connues qu'aux p points d'une grille fixée T_p. Le problème original est donc **semi-discrétisé**, et il faut se placer dans un sous-espace de H_μ, noté $H^p_\mu(T_p)$ ou plus simplement H^p_μ , dépendant simultanément de la grille et de la probabilité de référence. Nous avons montré qu'une **grille optimale** est obtenue lorsque les points échantillonnés sont des fractiles de μ. Dans ce cas, la variance généralisée (déterminant de la covariance) d'une famille quelconque de courbes cumulées de H^p_μ est minimale, d'où l'optimalité de la grille. La matrice de covariance d'une famille de courbes s'écrit alors $V \circ M^{-1}$, où V est la covariance ordinaire et M est la matrice de Gram de la base canonique de H^p_μ associée à T_p et K^μ. Dans le cas optimal, M^{-1} prend une forme particulièrement simple ([**P8***, p. 459] ; voir aussi la Section 1.4.1).

Concrètement, pour approcher $f = \frac{d\nu}{d\mu} = \Im^{-1}(F)$, il faut ensuite procéder à une deuxième étape de discrétisation, en choisissant un sous-espace de projection E_n engendré par n vecteurs linéairement indépendants de $\mathbf{L}^2_\mu$. Nous avons choisi celui engendré par les n premiers polynômes μ-orthogonaux (rangés par degré croissant), bien que leur calcul pose des problèmes numériques délicats. En effet, l'application associant aux K premiers moments $\left\{\int_a^b t^k d\mu(t),\ 0 \le k < K\right\}$ de la probabilité μ les coefficients des premiers polynômes μ -orthogonaux est très mal conditionnée [**Gaut82**]. Pour les calculer, nous avons utilisé l'algorithme de Chebyshev modifié, préconisé par

cet auteur. L'étape ultime consiste donc à résoudre le problème : trouver $f_n \in E_n$ tel que

$$\Im_p^n (f_n) = F_p \in H_\mu^p,$$

c'est-à-dire à calculer une inverse généralisée de la matrice I_p^n de cet opérateur. Cette matrice étant mal conditionnée, nous avons fait appel à deux méthodes de régularisation : la décomposition en valeurs singulières tronquée, et la régularisation de Tykhonov [**Han98**]. Dans cet ouvrage, Hansen montre le bon comportement de ces méthodes lorsqu'une condition technique (la condition discrète de Picard [**Han98**, p. 81-83]) est remplie ; il en est de même pour la principale méthode concurrente, la Validation Croisée généralisée. Malheureusement, la condition de Picard est typiquement associée aux opérateurs compacts alors que $\Im_p^n$ converge vers $\Im$, qui est unitaire ! Par conséquent, même si les méthodes de régularisation sont bien adaptées à la résolution du problème ci-dessus, aucun résultat théorique ne garantit vraiment la qualité de l'approximation de f_n obtenue par régularisation...

1.3. L'ACP de courbes granulométriques [P15*].

Il n'y a pas lieu cependant de trop s'alarmer de ce qui est dit ci-dessus : l'ACP a lieu dans $H_\mu^p(T_p)$, donc dans le cadre semi-discrétisé. La section précédente fournit surtout le cadre fonctionnel adapté à l'étude, via la métrique M^{-1} qui permet d'analyser le nuage des densités sans que l'on ait à dériver explicitement les courbes.

Il est fondamental de remarquer que la distance dans H_μ est donnée par la métrique du χ^2 centrée en μ :

$$\|\nu - \pi\|_{H_\mu}^2 = \int_a^b \left(\frac{d\nu}{d\mu} - \frac{d\pi}{d\mu} \right)^2 d\mu.$$

C'est pourquoi le choix de la loi uniforme $\mathcal{U}$ comme probabilité de référence est beaucoup moins naturel qu'on pourrait le penser ! On pourrait aussi bien prendre pour référence la moyenne $\overline{\nu}$ des courbes (ce qui mène directement à l'Analyse des Correspondances [**P15***]), ou une courbe particulière correspondant par exemple à un sédiment "source" transporté par le courant, ou même une fonction-poids associée à la physique du transport sédimentaire (c'est du reste ce que nous avons fait dans [**P15***]). Ces deux dernières options sont naturelles du point de vue du géologue.

Nous avons vu d'autre part que l'emploi d'une probabilité de référence uniforme donnera des résultats différents selon l'échelle utilisée. Pour prendre correctement en compte l'effet d'un changement d'échelle sur l'ACP, il faut observer que cela revient à **transporter les mesures** étudiées.

Deux échelles sont classiques en sédimentologie : l'échelle métrique usuelle ("Metric Units", MU) et l'échelle ϕ (logarithmique). Nous avons vu que les tailles échantillonnées sont généralement en progression géométrique. Par conséquent, dans le système ϕ, les classes granulométriques sont équilibrées (*i.e.* ont une largeur sensiblement constante) sur $[\log(a), \log(b)]$. Cette grille est optimale pour la probabilité de référence uniforme $\mathcal{U}$, autrement dit pour l'ACP classique. Cependant, les géologues pensent dans le système métrique mais, dans ce système, les classes ne sont plus du tout équilibrées ! La loi de référence transportée sur $[a, b]$ est en effet $\exp_* \mathcal{U}$, qui a pour densité usuelle $(\log(b) - \log(a))/y$, alors que la "densité usuelle" du sédiment ν est $f(y) := \frac{d \exp_* \nu}{d\mathcal{U}}(y)$. Nous avons donc dans le système MU un échantillonnage très serré des petites tailles, et une croissance linéaire de l'importance accordée aux particules en fonction de leur taille :

$$\frac{d \exp_* \nu}{d \exp_* \mathcal{U}} = \frac{d \exp_* \nu}{d\mathcal{U}} \frac{d\mathcal{U}}{d \exp_* \mathcal{U}} = f(y) \frac{y}{(\log(b) - \log(a))}.$$

.

Supposons par contre que nous puissions échantillonner de manière optimale les sédiments dans le système MU, et que nous passions dans le système ϕ. La loi de référence transportée sur $[\log(a), \log(b)]$ est $\log_* \mathcal{U}$, de densité $\exp(x)/(b-a)$, alors que le sédiment est classiquement

représenté par $g(x) := \frac{d\log_* \nu}{d\mathcal{U}}(x)$. Cette fois :

$$\frac{d\log_* \nu}{d\log_* \mathcal{U}} = \frac{d\log_* \nu}{d\mathcal{U}} \frac{d\mathcal{U}}{d\log_* \mathcal{U}} = g(x) \frac{(b-a)}{\exp(x)},$$

et l'on accorde beaucoup d'importance aux petites particules. Il est par conséquent impossible de dissocier l'échelle et la probabilité de référence : à chaque type d'analyse correspond un espace probabilisé de référence (e.p.r.) : $\{[a,b], \mathcal{B}([a,b]), \mu\}$. Un changement d'échelle $T : [a,b] \to [c,d]$ correspondra au transport de la mesure de référence vers le nouvel e.p.r. $\{[c,d], \mathcal{B}([c,d]), T_*\mu\}$. On a démontré dans **[P15*]** le théorème d'isométrie suivant.

THÉORÈME 1.1. *Soient ν et π deux mesures de H_μ, et T un changement de variable (un homéomorphisme). On a :*

$$\|\nu - \pi\|^2_{H_\mu} = \|T_*\nu - T_*\pi\|^2_{H_{T_*\mu}}.$$

Les résultats de l'ACP ne changent donc pas si les probabilités associées aux sédiments ainsi que la probabilité de référence sont transportées par le même homéomorphisme.

COROLLAIRE 1.2. *Soit F_μ la f.r. associée à μ. On a :*

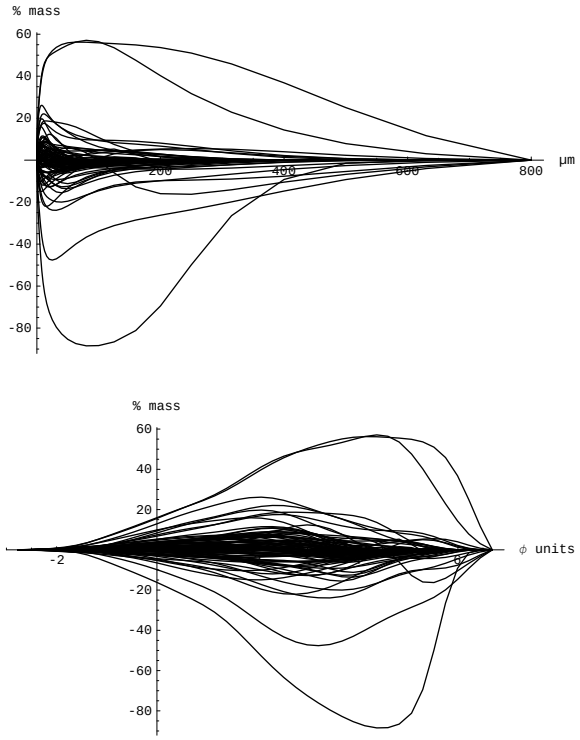
$$\|\nu - \pi\|^2_{H_\mu} = \|F_{\mu*}\nu - F_{\mu*}\pi\|^2_{H_\mathcal{U}},$$

car $F_{\mu*}\mu = \mathcal{U}$.

Autrement dit, toute ACP de mesures absolument continues peut se ramener au cas "standard " où l'e.p.r. est $\{[0,1], \mathcal{B}([0,1]), \mathcal{U}\}$ si F_μ est strictement croissante, ce qui est le cas ici car $\mu \approx \lambda$.

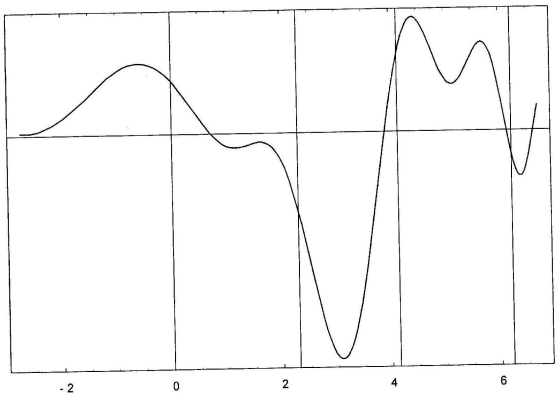
1.4. L'évolution sédimentologique de l'Etang de Berre (1992-1997).

FIG. 3.1. ***L'ensemble des différences $\Delta_s := F_s^{97} - F_s^{92}$, représentées dans les systèmes MU (panneau supérieur), et ϕ (panneau inférieur).***



Au cours de ces deux campagnes, une centaine de carottages ont été réalisés en des points fixes de l'étang, et les sédiments récoltés caractérisés de la même manière. L'analyse des courbes

FIG. 3.2. ***La densité moyenne des différences, $\frac{d\overline{\Delta}}{d\mathcal{U}}$, représentée dans le système ϕ. les bandes verticales séparent les types sédimentaires "canoniques" : colloïdes, argile, silt, sable et sable grossier.***



granulométriques de la première de ces campagnes a servi de support à [**P15***]. La méthode proposée permettant de traiter des mesures signées, nous allons ici analyser l'évolution ponctuelle de la granulométrie, c'est-à-dire les différences $\{\Delta_s,\ s = 1, \cdots, S\}$ entre courbes cumulées prélevées au même point (ou station).

Les courbes correspondant aux 91 stations communes aux deux campagnes sont représentées sur la figure 3.1. Les 42 tailles échantillonnées allant de 0.063 à $800\mu m$, nous avons pris comme support des mesures dans le système MU l'intervalle $[a, b] := [0.05\mu m, 1260\mu m]$. L'aspect des courbes dépend fortement de l'échelle utilisée, que nous allons fixer comme étant ϕ.

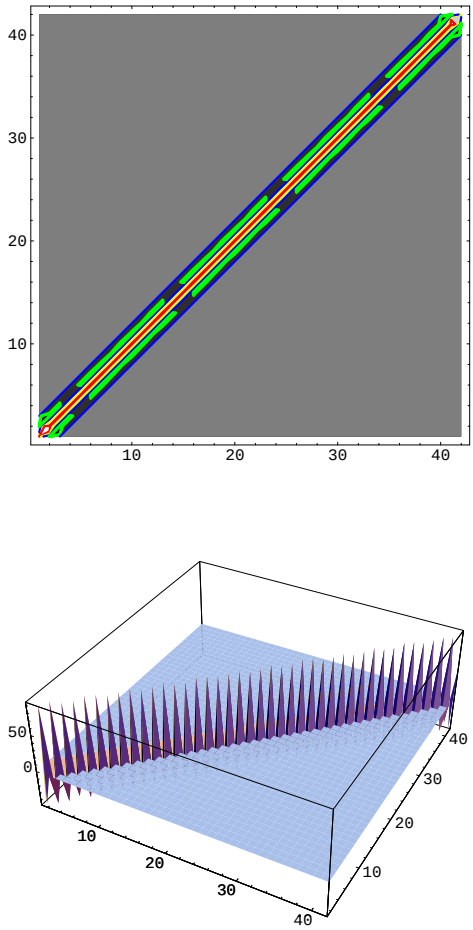
La première caractéristique à considérer est la différence moyenne $\overline{\Delta}$, dont la dérivée de Radon-Nikodym par rapport à la distribution uniforme (approchée dans H_ϕ) est représentée sur la figure 3.2. Le graphe de cette fonction montre que de 1992 à 1997 la granulométrie de l'étang s'est appauvrie en silt, avec un enrichissement corrélatif en sable fin et colloïdes.

L'objet de l'ACP consiste à résumer la structure des écarts centrés $\left\{\Delta_s - \overline{\Delta},\ s = 1, \cdots, S\right\}$. Nous allons dans un premier temps détailler les résultats de l'analyse dans l'espace H_ϕ associé à l'e.p.r. $\{[\alpha, \beta], \mathcal{B}([\alpha, \beta]), \mathcal{U}\}$, avec $\alpha = \log(a)$ et $\beta = \log(b)$. Ensuite, nous donnerons les résultats de l'analyse dans l'espace H_{MU} correspondant à l'e.p.r. $\{[\alpha, \beta], \mathcal{B}([\alpha, \beta]), \log_* \mathcal{U}\}$, alias $\{[a, b], \mathcal{B}([a, b]), \mathcal{U}\}$.

1.4.1. *L'ACP dans H_ϕ.*

La grille utilisée étant quasi-optimale, la métrique M_ϕ^{-1} est très proche de la matrice théorique correspondant à la grille optimale de taille $p = 42$ [**P8***, p. 459]. Les éléments de cette matrice étant exclusivement à valeurs dans $\{-42, 0, 42, 84\}$, nous avons utilisé ces valeurs privilégiées pour la représenter dans la figure 3.3, qui montre qu'elle accorde pratiquement le même poids à toutes les classes granulométriques.

FIG. 3.3. ***Représentations de M_ϕ^{-1}. Panneau supérieur : tracé de contour des valeurs -42 (bleu), 0 (vert), 42 (jaune) et 84 (rouge). La zone grise correspond à des valeurs proches de zéro. Panneau inférieur : tracé tridimensionnel.***



L'essentiel de la variance (63%) est résumé par la première composante, $CP1_\phi$, représentée sur la figure 3.4. La première partie de cette figure contient un histogramme de cette composante. Il est dit "optimal", car il minimise l'Erreur Quadratique Intégrée asymptotique (dans le cas Gaussien) ; sa largeur de classe h est liée à l'écart-type de $CP1_\phi$ et au nombre n des stations par la relation $h_n = 3.5\sigma_n\, n^{-1/3}$ [**BL87**, p. 150-153]. On observe que la distribution de $CP1_\phi$ est plutôt lacunaire, la moitié des 14 classes de l'histogramme étant vides. Cette composante est étroitement associée à trois stations extrêmes, les contributions des autres étant très faibles.

La nature de $CP1_\phi$ peut être expliquée par la fonction propre associée [**P15***, Section 3.1], notée $\psi_\phi^{\mathcal{U}}$, qui est **l'approximation** de la densité $d\Psi_\phi/d\mathcal{U}$ obtenue en inversant l'opérateur $\Im_p^n$ au moyen d'une méthode de régularisation. Ici Ψ_ϕ désigne le premier vecteur propre de $V \circ M_\phi^{-1}$. La fonction $\psi_\phi^{\mathcal{U}}$ est représentée dans le panneau inférieur gauche de la figure 3.5. Dans le panneau inférieur droit est représentée la densité de cette mesure signée, relative à la mesure de Lebesgue (identique à la précédente, à une constante multiplicative près). Sur le panneau supérieur droit est tracée en rouge l'intégrale $\int_a^s \psi_\phi^{\mathcal{U}}(t)\, d\mathcal{U}(t)$, à laquelle est superposé le vecteur Ψ_ϕ (points noirs). On peut constater que l'approximation est excellente ; les résidus sont représentés sur le panneau supérieur gauche. Le graphe de $\psi_\phi^{\mathcal{U}}$ montre très clairement que $CP1_\phi$ oppose le sable fin (et éventuellement le silt) au sable grossier ; le reste des catégories granulométriques ne joue aucun rôle notable.

FIG. 3.4. ***Panneau supérieur : histogramme de la première composante de l'ACP dans la métrique*** M_ϕ^{-1}, $CP1_\phi$. ***Panneau inférieur : cartographie de*** $CP1_\phi$ ***; la couleur des disques correspond aux classes de l'histogramme.***

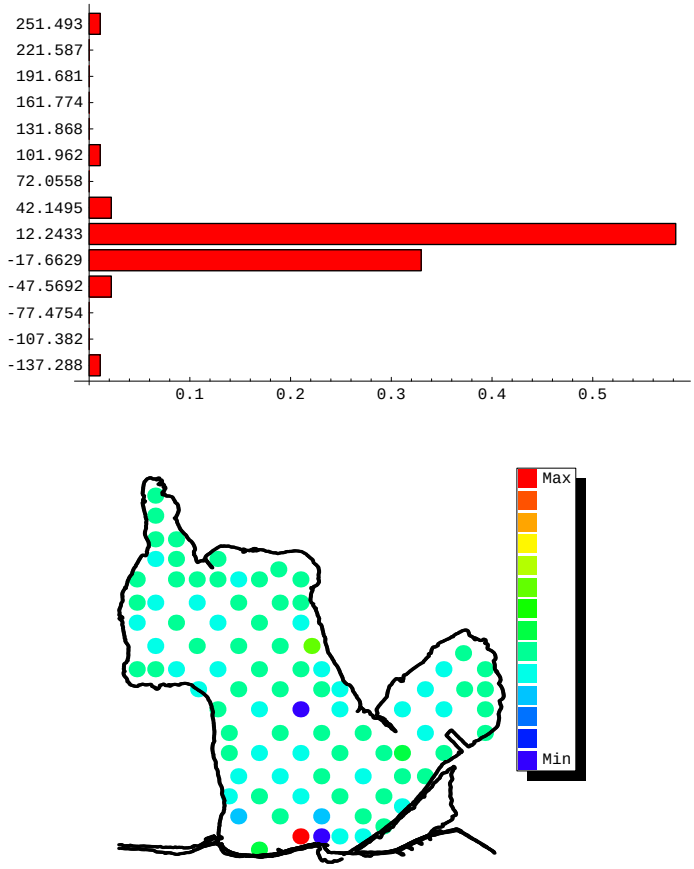
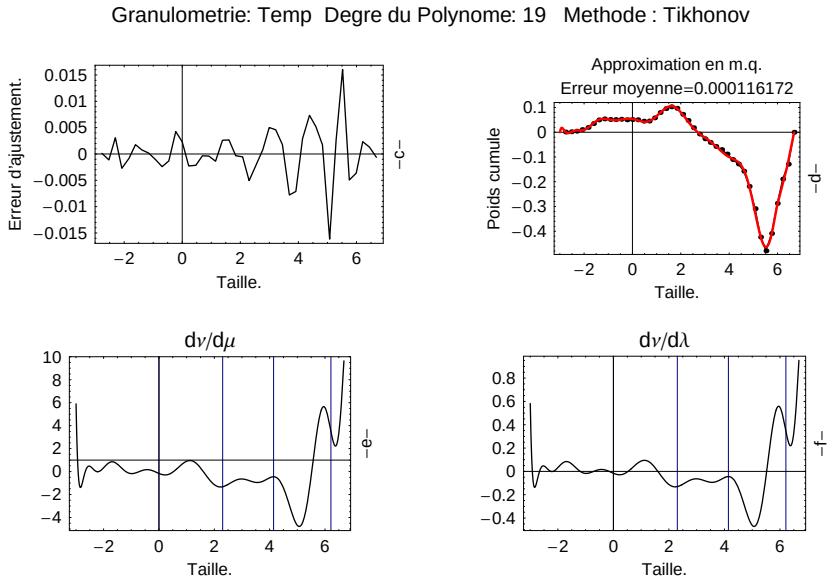


FIG. 3.5. ***La première fonction propre de l'ACP dans la métrique*** M_ϕ^{-1} ***(commentaires dans le texte).***



1.4.2. *L'ACP dans H_{MU}.*

La structure de la métrique de cet espace (représentée sur la figure 3.6) est bien différente de la précédente. La matrice associée est creuse elle aussi : 93% de ses éléments sont de module inférieur à 10^{-6}. Les premières valeurs situées sur la diagonale sont très fortes relativement à celles de M_ϕ^{-1} . Cette métrique accorde donc un poids beaucoup plus important que la précédente aux fractions les plus fines.

L'essentiel de l'information (environ 83% de la variance) est résumé par la première composante $CP1_{MU}$ représentée sur la figure 3.7. On notera que l'histogramme optimal nécessite moins de classes (10) que dans le cas précédent, et qu'une seule d'entre elles est vide. On retrouve bien sur la carte du panneau inférieur les trois stations principales de l'analyse précédente, mais d'autre stations apparaissent aussi comme importantes.

Nous pouvons ici aussi expliquer la nature de la composante par la fonction propre $\psi_{MU}^{\log_* \mathcal{U}}$ approchant $d\Psi_{MU}/d\log_* \mathcal{U}$. Le panneau inférieur gauche de la figure 3.8 montre que $CP1_{MU}$ dépend uniquement de la partie la plus fine des sédiments : les colloïdes et, marginalement, l'argile. Le graphe de la densité usuelle $d\Psi_{MU}/d\lambda$ montre cependant que les sables jouent aussi un rôle, mais passif (sur cette figure, Ψ_{MU} est "vu" dans le système ϕ), de manière tout à fait analogue à ce qui se produit habituellement pour des variables supplémentaires. Cela explique la similitude de la carte de $CP1_\phi$ avec celle de $CP1_{MU}$: seules les parties extrêmes de la granulométrie varient, en sens opposé ; la partie centrale est stable à l'échelle de l'étang.

FIG. 3.6. ***Représentation de M_{MU}^{-1}. Panneau supérieur : zones noires (resp. claires) : valeurs faibles (resp. positives fortes) ; le contour vert correspond aux zéros. Panneau inférieur : tracé tridimensionnel.***

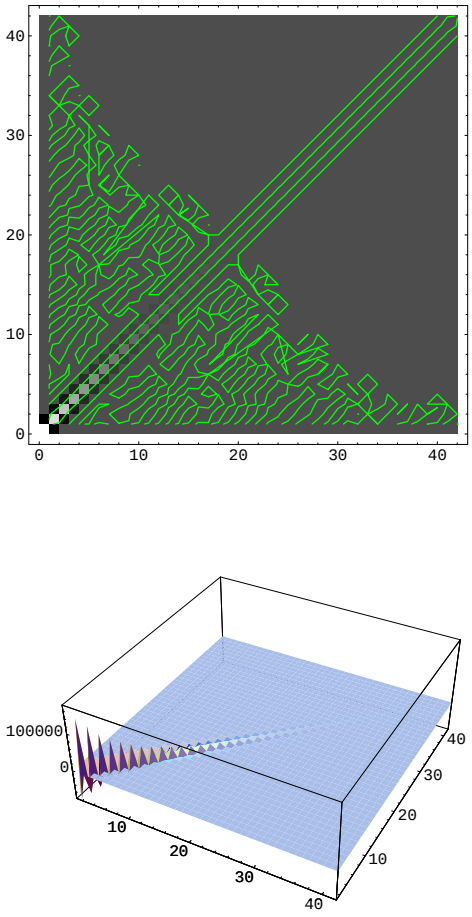


FIG. 3.7. ***Première composante de l'ACP dans la métrique*** M_{MU}^{-1}***. La légende est la même que dans 3.4.***

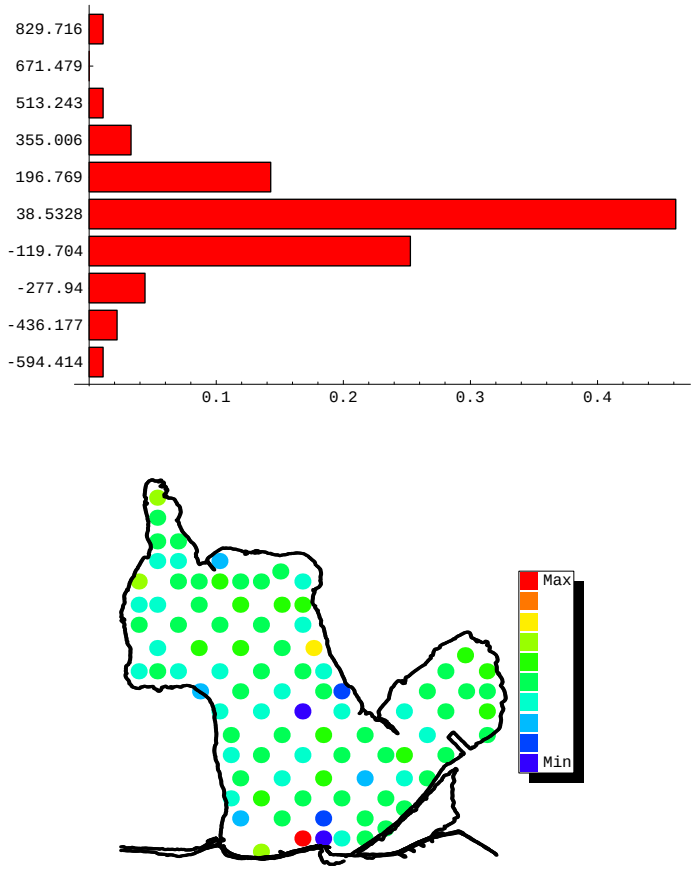
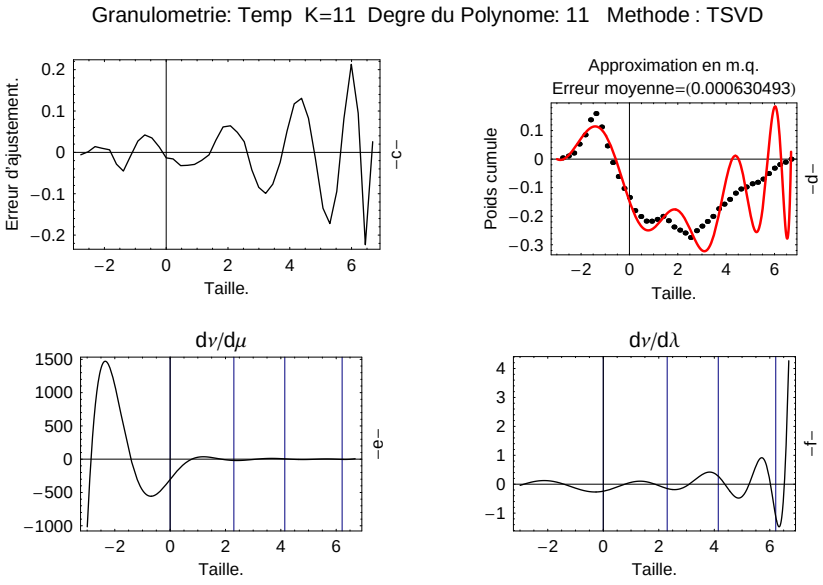


FIG. 3.8. ***La première fonction propre de l'ACP dans la métrique*** M_{MU}^{-1} ***(commentaires dans le texte).***



1.4.3. *Discussion.*

L'analyse est intellectuellement plus satisfaisante dans H_{MU} que dans H_ϕ car, comme dans [**P15***], les composantes principales y sont moins bruitées et le cadre fonctionnel se justifie mieux du point de vue exploratoire : il est plus naturel de raisonner dans le système métrique MU. Par contre, la différence entre les cartes obtenues n'est guère spectaculaire, alors qu'elle l'était dans [**P15***]. De plus, on remarquera que la qualité de l'approximation de l'intégrale :

$$\int_a^s \psi_\phi^{\log_* \mathcal{U}}(t)\, d(\log_* \mathcal{U})(t) = \frac{1}{(b-a)} \int_a^s \psi_\phi^{\log_* \mathcal{U}}(t)\, \exp(t)\, dt$$

est très mauvaise pour la partie grossière du sédiment (silt, sables), comme on le voit dans les panneaux supérieurs de la figure 3.8. Cela se comprend dans la mesure où la pondération de ces fractions dans l'ACP fait qu'elles sont pratiquement passives. Cependant, on ne peut pleinement se satisfaire de cette explication, car l'analyse dans H_{MU} pose plusieurs problèmes pratiques.

La source de ces difficultés réside dans le fait que la grille d'échantillonnage utilisée est très loin d'être optimale. Par conséquent, le calcul de la métrique M_{MU}^{-1} peut être entaché d'erreurs se répercutant sur l'analyse, d'autant plus que le calcul des polynômes orthogonaux associés à l'e.p.r. $\{[\alpha, \beta], \mathcal{B}([\alpha, \beta]), \log_* \mathcal{U}\}$ pose aussi des problèmes numériques : nous ne sommes arrivés à en calculer que douze avec une précision suffisante. Une densité ne peut donc ici être approchée que dans le "petit" sous espace E_{12} de $\mathbf{L}^2_\mu$. Est-il "suffisant", autrement dit, un objet d'intérêt de H_{MU} (Ψ_{MU} par exemple) appartient-il généralement à l'espace $\Im_p^{12}(E_{12})$? Bien évidemment non ! Le remède à ces difficultés se trouve dans le Corollaire 1.2 : il suffirait d'interpoler les courbes granulométriques aux points d'une grille μ-optimale bien choisie. Le problème du calcul de la métrique serait alors résolu. Il serait également inutile de calculer les polynômes orthogonaux : on aurait seulement à choisir entre la base de Fourier et les polynômes de Legendre. C'est ce que nous ferons dans la prochaine version du programme.

Cependant, les résultats ci-dessus mettent en évidence une insuffisance méthodologique d'une autre nature : l'analyse proposée ne tient aucun compte de la nature spatiale des données. Cela n'est pas gênant lorsque les composantes sont naturellement spatialement organisées, ce qui était le cas dans [**P15***]. Mais ici, les cartes factorielles représentées sur les figures 3.7 et 3.4 ne montrent aucune structure claire : il semble que l'évolution granulométrique soit essentiellement résumée par la différence moyenne représentée sur la figure 3.2, et quelques singularités ponctuelles peut-être sans signification (erreur d'échantillonnage, passage d'un bateau, ...). Par construction, l'ACP ne produit pas des composantes spatialement décorrélées, mais il existe une abondante littérature à ce sujet. L'approche classique vient des géostatistiques, où nombre d'auteurs ont adapté le krigeage au cas multivarié (voir par exemple [**Wa98, AE01**]). La méthode consiste en gros à approcher la matrice de covariance par une somme de K (à déterminer) matrices dites de corégionalisation, associées à des structures spatialement décorrélées, et attachées chacune à un variogramme. La diagonalisation séparée de ces K matrices fournit un ensemble de variables non corrélées dans les deux sens (ordinaire et spatial). Mais cette méthode est gourmande en temps de calcul, et dépend des modèles (arbitraires) de variogramme choisis [**AE01**]. Sicard *et al.* [**SSN02**] ont par conséquent proposé une méthode plus économique, basée elle aussi sur le krigeage. Elle consiste, à partir du premier vecteur propre de l'ACP, à le modifier itérativement de façon à ce que la composante obtenue tende à être conforme au modèle de variogramme choisi. Cette méthode permet même de traiter des données plus complexes (un cube de données), en utilisant STATIS. Enfin, Krzanowski & Bailey [**KB07**] proposent une autre méthode, consistant à rechercher des composantes ponctuellement décorrélées dont la fonction d'autocorrélation spatiale est conforme à un modèle, exponentiel par exemple.

2. Une application en écologie des populations

Les probabilités de référence utilisées ci-dessus se ramènent à la loi uniforme, modulo un changement d'échelle. Mais la probabilité de référence permet aussi d'orienter l'analyse en introduisant des connaissances a priori. Par exemple, dans [**P15***], $\mu(x)$ était proportionnelle à la tension de fond, c'est-à-dire à l'énergie nécessaire pour mettre en suspension une particule de

taille x. Les résultats de l'ACP dépendaient donc de ce que l'on connaît des lois du transport sédimentaire. Un autre exemple d'application [**P14**] concerne l'écologie benthique.

Supposons que l'on s'intéresse au comportement temporel de S espèces et que l'on dispose, pour chaque espèce e, d'un tableau

$$\{C_e(r,t), 1 \leq t \leq T, 1 \leq r \leq R\}$$

recensant le nombre de membres de e trouvés dans le réplicat r au temps t (à chaque instant de prélèvement, R réplicats ont été échantillonnés). Le comportement collectif (grégaire, « aléatoire », ou répulsif) à l'instant $t \in \{1, \cdots, T\}$ de e est classiquement caractérisé par l'indice de dispersion :

$$I_e(t) = \frac{\sum_{r=1}^{R} \left(C_s(r,t) - \overline{C_s}(t)\right)^2}{\overline{C_s}(t)}.$$

Il est établi que si l'espèce e a un comportement « aléatoire » (en fait : poissonnien), I_e suivra asymptotiquement à chaque instant la loi χ^2_{R-1} , lorsque le nombre total de représentants de e tend vers l'infini. La fonction de répartition empirique de I_e caractérise donc le comportement collectif de e, et il est naturel de choisir χ^2_R comme probabilité de référence pour étudier les espèces du point de vue de leur comportement. C'est ce qui a été fait dans [**P14**].

CHAPITRE 4

La turbulence et le plancton

Sommaire

L'esprit n'use de sa faculté créatrice que quand l'expérience lui en impose la nécessité. [**Po43**, p.43]

Au coeur de ce travail se trouve le "paradoxe du plancton" que l'on peut résumer ainsi : pourquoi y-a-t-il une telle biodiversité dans les océans, alors qu'il s'y trouve si peu de ressources ? Pour y répondre (sur le registre numérique), nous projetons de faire intervenir la turbulence dans les modèles de croissance du phytoplancton, ce qui nous obligera à simuler de manière convaincante un phénomène dont la modélisation (sur une base physique) est toujours un défi pour les spécialistes. Ce chapitre mêlera donc l'Ecologie Théorique, l'Océanographie Physique et la Statistique des Processus. Cette incursion en territoire étranger n'est pas sans péril, car notre approche de la turbulence n'est pas celle des physiciens : elle n'a aucune ambition explicative, et ne vise qu'à produire des entrées réalistes pour les modèles numériques.

1. Le paradoxe du plancton

Les modèles usuels de compétition entre deux populations phytoplanctoniques avec un sel nutritif limitant dans un milieu homogène prédisent l'extinction d'une espèce, pour presque tous les points de l'espace des paramètres [**Ha60**] (pour ce qui concerne les milieux contrôlés, voir [**SW95**]). Ce résultat mathématique est confirmé par de nombreuses expériences en culture, à commencer par le travail de Gause [**Gaus35**], qui ont donné naissance au "Principe d'exclusion compétitive". Or, cela est en contradiction avec ce que l'on observe dans la nature (par exemple dans le cas du phytoplancton), où de très nombreuses espèces concurrentes coexistent : c'est le **paradoxe du plancton**. Nous projetons de le résoudre en répondant à la question suivante : la turbulence, en favorisant le contact entre les cellules et le nutriment [**Lag06**] de même qu'elle affecte les rencontres entre prédateurs et proies [**SSL01**], permet-elle d'expliquer la biodiversité du phytoplancton ?

Dans le cadre d'un modèle déterministe, Poggiale *et al.* [**P18**] ont répondu à une question similaire, en montrant que la dynamique d'une population répartie sur deux sites peut grandement différer de celle d'un milieu homogène. Plus précisément, dans les conditions du modèle, l'hétérogénéité spatiale d'un milieu augmente mécaniquement sa productivité. La stratification du milieu marin (en salinité, température, oxygénation, illumination, *etc.*) correspond bien au type d'hétérogénéité pris en compte par ce modèle, dont les auteurs ont de plus montré qu'il n'est pas affecté par l'ajout d'un bruit blanc "environnemental". Mais le milieu marin n'est pas seulement stratifié : il est en permanence brassé par la turbulence. Une cellule planctonique (surtout passive, ce qui est le cas du phytoplancton) divague donc dans un milieu hétérogène, dont une des caractéristiques est l'intermittence [**Man89**]. Celle-ci est-elle responsable du paradoxe, en favorisant l'apport de nutriments aux cellules [**Lag06**] ? L'adoption de cette hypothèse mène à

incorporer la turbulence (au travers de l'Energie Cinétique Turbulente : ECT) dans le modèle utilisé.

La structure d'un modèle simple de croissance d'une espèce phytoplanctonique en milieu turbulent est du type :

$$\begin{array}{l} \frac{dX}{dt} = -\frac{aX}{b+X}Y + \varepsilon(t) \\ \frac{dY}{dt} = \left(e\frac{aX}{b+X} - m\right)Y \end{array} \tag{4.1}$$

où X désigne la concentration en nutriment, Y la densité du phytoplancton, m son taux de mortalité, a la vitesse maximale spécifique d'absorption du sel nutritif, b la constante de demi-saturation de la cinétique d'absorption, et e l'efficacité de conversion. Le terme aléatoire $\varepsilon(t)$ désigne ici l'apport de sel nutritif, supposé proportionnel à l'ECT ; nous confondrons ici purement et simplement ces deux quantités.

Etant donnée la nature de la turbulence, une série chronologique d'ECT ne peut être considérée comme une suite figée de valeurs, mais plutôt comme une **trajectoire** particulière d'un processus aléatoire. Notre objectif sera donc de proposer un modèle statistique de l'ECT, tel que que l'on puisse facilement estimer les paramètres des trois trajectoires dont nous disposons, afin d'en simuler un grand nombre. L'amplitude de l'écart entre les différentes solutions numériques de (4.1), associées à différentes trajectoires de l'ECT, et la solution de l'équation déterministe (privée de $\varepsilon(t)$) devrait fournir une réponse à notre question. Pour cela, il est nécessaire que l'équation différentielle stochastique (4.1) soit intégrable. Il faudrait donc au minimum qu'une expression du type $\int U_t\,\varepsilon(t)\,dt$ ait un sens, lorsque U_t désigne un des processus intervenant dans cette équation. Notre pseudo-ECT doit donc être un modèle réaliste, mais pas trop "sauvage", de l'ECT. Le moment est donc venu d'aborder succinctement un sujet vaste mais incontournable : la modélisation de la turbulence.

2. La turbulence en deux mots (ou presque)

Voici ce que dit Uriel Frisch [**Fr02**] pour souligner la complexité du phénomène :

> *... le sujet est très interdisciplinaire et touche...à la physique, à la mécanique des fluides, à la météorologie et à l'astrophysique. Après une brève introduction, je vous dirai deux mots de la formulation du problème, puis je vous parlerai de transition, de chaos, d'effet papillon, de mouvement brownien, de chou-fleur et enfin du million de dollars que M. Clay nous a promis.*

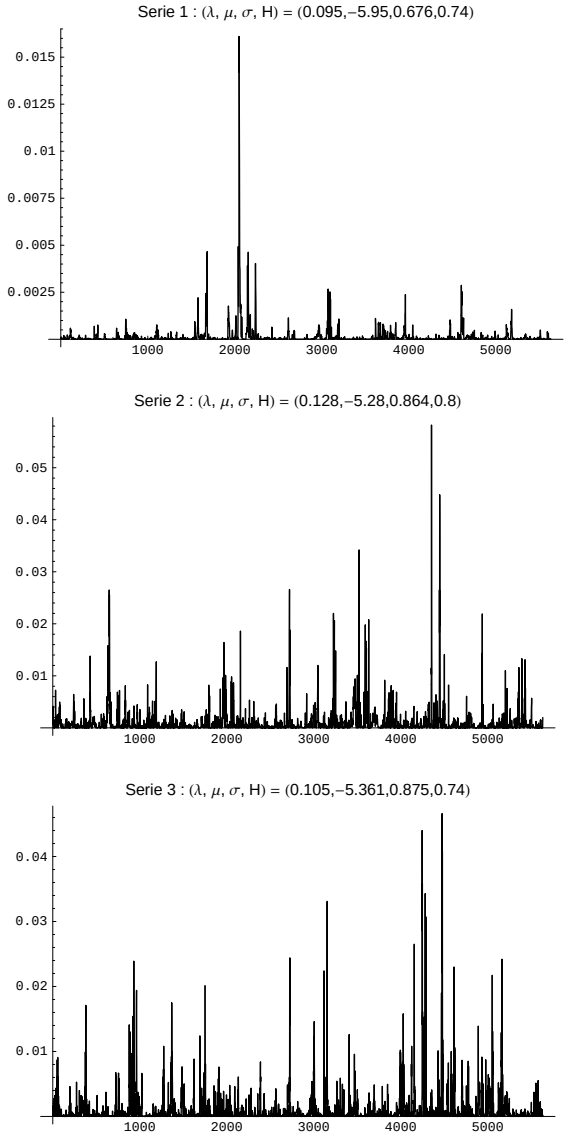
Plutôt que de céder à un penchant naturel en nous intéressant au million de dollars, nous allons nous pencher sur l'humble chou-fleur, métaphore de l'auto-similarité et des fractales.

Avant l'apparition des fractales, régnait en mécanique des fluides la turbulence homogène de Kolmogorov (1941), c'est-à-dire le modèle de la cascade multiplicative homogène : un tourbillon, occupant à l'instant t tout le volume du fluide, se divise à l'instant $t+1$ en k tourbillons de même volume, se partageant une fraction de l'espace. Mais cette théorie, à peu près satisfaisante pour les ingénieurs, ne l'était pas pour les physiciens, car elle ne respectait pas la nature intermittente du phénomène. Mandelbrot [**Man89**, chapitres VIII et IX] résolut ce problème en introduisant les mesures multifractales, dont l'archétype est la très cantorienne mesure fractale multinomiale : l'énergie se concentre sur un support extrêmement morcelé, de mesure de Lebesgue asymptotiquement nulle (sa distribution est une mesure singulière).

Le formalisme multifractal est utilisé dans la modélisation d'une multitude de phénomènes (en physique, finance, imagerie, trafic internet, circulation routière, etc.), mais l'approche la plus fréquente en Météorologie et en Océanologie est celle des multifractales universelles [**SL02, SSL01**] de Schertzer & Lovejoy. Ce modèle de cascade multiplicative ne dépend que de trois paramètres (pour un exposé assez clair, voir [**TL00, LS07**]) : C_1, caractérisant la dimension fractale moyenne des trajectoires, et les deux paramètres du processus multiplicatif (log-stable [**ST94**]) : son paramètre de stabilité α, et un paramètre d'échelle H (coïncidant avec l'exposant de Hurst dans le cas monofractal). Il est à noter que le paramètre de dissymétrie du processus est fixé à la valeur $\beta = -1$, de façon à ce que ses moments existent [**SL02**].

3. Estimation des paramètres et simulations

FIG. 4.1. ***Les trois trajectoires d'ECT disponibles.***



Mais l'estimation des paramètres d'une multifractale universelle (et des multifractales en général) est difficile et, surtout, quel sens a une intégrale du type $\int U_t\, \varepsilon(t)\, dt$, lorsque $\varepsilon(t)$ est un tel processus ? Nous avons donc suivi une autre voie, en modélisant l'ECT par un bruit Gaussien fractionnaire (bGf) convenablement transformé. Les paramètres du processus sont faciles à estimer, et la simulation d'un bGf est classique. Le modèle proposé n'a certes aucun sens physique, mais rappelons que notre seul but est de générer un nombre arbitraire de trajectoires possédant une structure statistique semblable à celle des données représentées sur la figure 4.1. Notons que ces séries chronologiques sont toutes de longueur raisonnable (5632).

Chaque trajectoire expérimentale $\check{\varepsilon}(t)$ est une réalisation du processus $\varepsilon(t)$, et nous permet d'estimer les quatre paramètres de celui-ci :

- son paramètre de Box-Cox λ , estimé par maximum de vraisemblance
- la moyenne μ et l'écart-type σ de la chronique transformée ζ
  $\check{\zeta}(t) := \left(\check{\varepsilon}(t)^{\lambda} - 1\right)/\lambda$
- le paramètre de Hurst H du processus stationnaire et Gaussien $\zeta(t)$.

Le seul de ces paramètres dont l'estimation pose problème est l'exposant de Hurst. Considérons l'intégrale stochastique :

$$\Omega(t) := \int_0^t \zeta(s)\, ds.$$

Lorsque $\zeta(t)$ est un bGf de paramètre H centré réduit, Ω est le mouvement Brownien fractionnaire standard de même paramètre, noté B_H. La propriété saillante de B_H est son auto-affinité, c'est-à-dire que l'on a [**Man97**, p.58] :

$$B_H(t+\Delta t) - B_H(t) \sim (\Delta t)^H.$$

Par conséquent, H est généralement estimé à partir de la pente de la régression linéaire entre le logarithme du "pas d'échantillonnage" Δt et celui d'une statistique $\Sigma(\Delta t)$ portant sur l'incrément $\Omega(t+\Delta t)-\Omega(t)$, que le processus soit Gaussien ou non. La statistique en question peut être bâtie sur différentes caractéristiques de $\zeta(t)$: fonction d'autocorrélation, variogramme, variance, R/S de Hurst [**Ber94, Man97**]. Le point intéressant est que lorsque $\zeta(t)$ est échantillonnée avec un pas constant δ, la durée de la chronique est $T = N\,\delta$, alors que $\Delta t = k\,\delta$. L'estimation de H est donc bâtie sur l'ensemble $\{(k, \Sigma(k)) : k \in D(N)\}$ où $D(N)$ désigne l'ensemble des diviseurs non triviaux de N. Rien n'est donc possible si N est premier ! En partant de cette simple constatation, nous avons pu montrer dans [**P16***] que l'estimation de H par une méthode classique (tracé de la variance des moyennes locales) est toujours améliorée lorsque l'on exécute la procédure standard sur des sous-chroniques bien construites. La méthode consiste à tirer au hasard de manière appropriée le début et la fin de chaque sous-chronique, et de n'utiliser pour l'estimation que les sous-chroniques dont la longueur possède le plus possible de diviseurs. Cela améliore la qualité des régressions, donc celle de l'estimateur.

Le deuxième point abordé dans [**P16***] est le fait que, mécaniquement, les valeurs de $\Sigma(k)$ sont moins fiables lorsque k est "grand" car, dans ce cas, le nombre de fenêtres est "petit". Nous avons donc eu l'idée d'améliorer l'estimation de H en remplaçant la régression linéaire classique par une procédure de régression biphasée [**Lar92**]. On estime alors H à partir de la pente de la droite de régression obtenue pour les fenêtres étroites ; la valeur critique de la largeur k est déterminée automatiquement par l'algorithme de Larson. Nous avons introduit la même procédure de régression biphasée pour estimer H avec la méthode du périodogramme (en nous focalisant cette fois sur les basses fréquences), mais avec beaucoup moins de bonheur.

En analysant la chronique de l'Oscillation Nord-Atlantique, nous avons pu montrer que la méthode permet de mettre en évidence la non-stationnarité éventuelle du processus étudié. On pourrait facilement l'adapter à d'autres estimateurs, faisant appel à d'autres statistiques $\Sigma(k)$ (le R/S de Hurst, par exemple).

3.1. Retour sur l'équation (4.1).

On peut voir sur la figure 4.1 que, malgré les apparences, nos trois trajectoires d'ECT on à peu près les mêmes paramètres, et que le processus est persistant, avec un exposant de Hurst proche de 0.75. Remarquant que λ est toujours proche de 0.1, nous poserons :

$$\zeta(t) := 10\left(\varepsilon(t)^{0.1} - 1\right) \tag{4.2}$$

ce qui est très commode car, si l'on admet que $\zeta(t) = \frac{dB_H}{dt}(t)$, il vient :

$$\begin{aligned} \int U_t\, \varepsilon(t)\, dt &= \int U_t \left(\tfrac{\zeta(t)}{10} + 1\right)^{10} dt \\ &\approx \lim_{\delta\to 0} \int U_t \left(\tfrac{B_H(t+\delta) - B_H(t)}{10\,\delta} + 1\right)^{10} dt. \end{aligned}$$

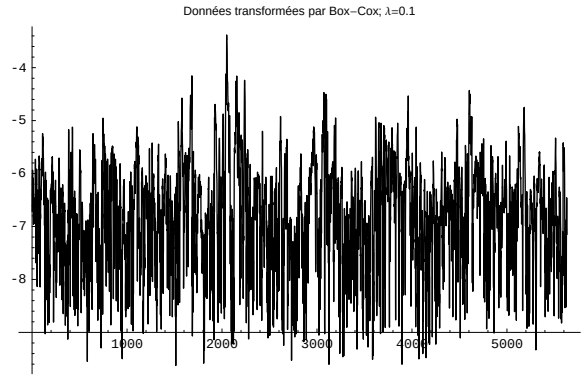
Or, Gradinaru et Nourdin [**GN03**] montrent que, si $m = 2n$:

$$\int U_t \left(\frac{B_H(t+\delta) - B_H(t)}{\delta}\right)^m dt \approx \delta^{2nH-1} \frac{(2n)!}{2^n n!} \int U_t\, dt$$

et que, si m est impair, cette intégrale stochastique converge presque sûrement vers zéro. Tout cela laisse donc supposer que les solutions numériques de l'équation (4.1) seront convergentes. Le problème est cependant compliqué par le fait que nous avons affaire à un système différentiel...

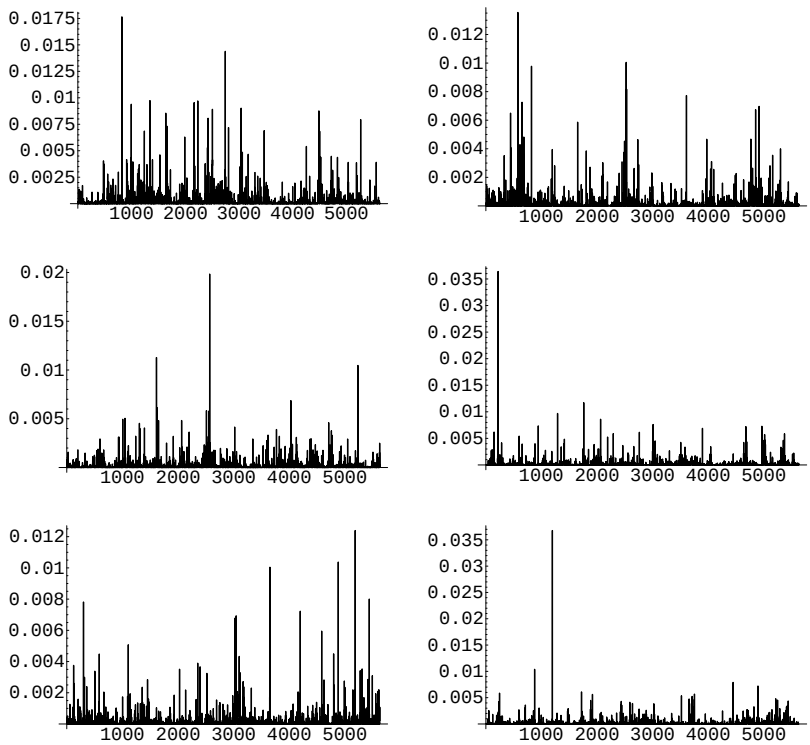
3.2. Un exemple de simulation.

FIG. 4.2. ***La première trajectoire après transformation.***



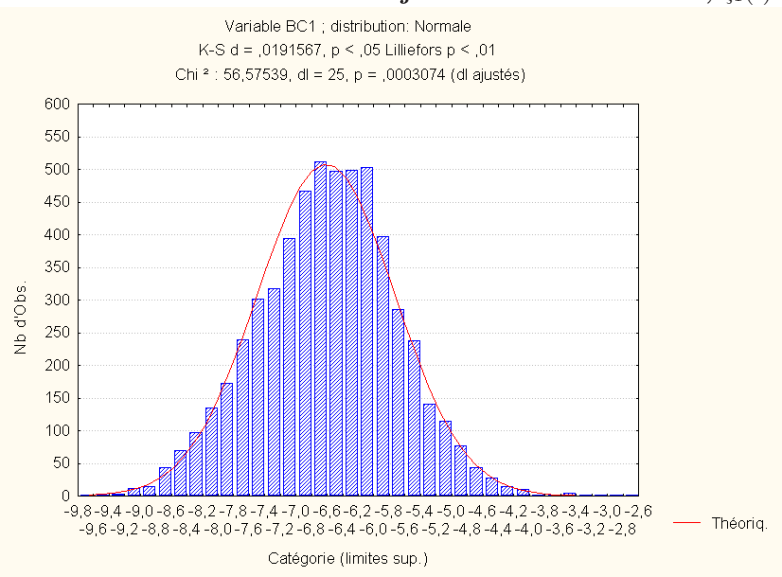
La réalisation de $\zeta(t)$ représentée sur la figure 4.2 est la transformée par (4.2) de la première trajectoire de la figure 4.1. Ce processus semble bien stationnaire et Gaussien (mais voir la Section 3.3, pour un bémol). Pour simuler des trajectoires similaires, nous avons utilisé la méthode de Dietrich & Newsam [**DN97**]. Elle est rapide, exacte [**BL03**], et permet de simuler tout processus stationnaire Gaussien de fonction de covariance connue. La simulation de trajectoires de l'ECT consiste donc à appliquer la transformation inverse de (4.2) à des réalisations du bGf de paramètres $(\mu = -5.95, \sigma = 0.676, H = 0.74)$. Un petit nombre de réalisations sont représentées sur la figure 4.3. Noter la similitude avec les signaux de la figure 4.1.

FIG. 4.3. ***Six simulations de la trajectoire N°1.***



3.3. Extensions non-multifractales.

FIG. 4.4. ***Distribution de la trajectoire N°1 transformée,*** $\check{\zeta}_1(t)$.



Bien que les séries chronologiques transformées aient une distribution très proche de la normalité, elles ne "passent" pas les tests non-paramétriques classiques au seuil de 5%, comme on peut le voir sur la figure 4.4. On peut constater que pour les trois séries transformées, le modèle Gaussien surestime la fréquence des très petites valeurs (crée de l'intermittence, en quelque sorte), et sous-estime la fréquence des très fortes valeurs. Si le premier point n'est guère gênant (peu importe pratiquement que $\zeta(t)$ soit très petit ou très très petit), l'existence de fortes (mais rares !) bouffées de turbulence pourrait jouer un rôle important dans la résolution numérique du système (4.1).

Que faire ? Dans un premier temps, nous avons simulé des trajectoires "quasi-Gaussiennes" : leur fonction de corrélation est celle d'un bGf, mais leur distribution marginale possède des coefficients d'aplatissement et de dissymétrie (et éventuellement des termes d'ordre plus élevé) non nuls. Ce pis-aller permet d'obtenir des pics plus intenses sur les trajectoires de $\zeta(t)$, mais la perte du modèle Gaussien ne nous permet théoriquement plus d'utiliser les résultats de la Section 3.1 ! Une autre manière canonique de s'écarter de la normalité est de faire appel aux lois stables. Nous l'avons fait de deux façons :

(1) en approchant la loi de ζ par une loi stable

(2) en approchant la loi de ε par une loi log-stable.

La première approximation est semblable au développement de Gram-Charlier proposé ci-dessus, avec le même inconvénient majeur : on ne peut plus rigoureusement appliquer les résultats de **[GN03]**. Pour estimer les paramètres, nous avons utilisé la version *Mathematica* du programme STABLE de J. P. Nolan (`http://www.robustanalysis.com/`). Les estimations, effectuées par maximum de vraisemblance **[No01]**, sont ici encore très proches les unes des autres, confirmant le fait que les données sont sans doute générées par un même processus. Mais le fait principal est que, dans tous les cas, la loi estimée est normale (le paramètre de stabilité est égal à deux). Ceci dit, le test d'adéquation de Kolmogorov-Smirnov conduit à rejeter l'hypothèse de normalité dans le cas des deux premières trajectoires (pour $\check{\zeta}_3$, le niveau de signification vaut 0.095, et la normalité serait admissible).

La deuxième approximation est meilleure pour les deux premières trajectoires, mais moins bonne pour la troisième ; le niveau de signification est toujours inférieur à 5%, ce qui rend l'hypothèse de stabilité douteuse. Si ce modèle était néanmoins adopté, il est possible (mais délicat, voir **[ST94**, Section 7.11]) de construire des trajectoires du processus $\zeta(t) := \log(\varepsilon(t))$. Mais quel serait le sens de l'expression $\int U_t\, \varepsilon(t)\, dt$? Même l'existence d'intégrales de chemin du type $\int Z_t\, \zeta(t)\, dt$ n'est pas immédiate dans ce cadre **[ST94**, Section 11.2]...

CHAPITRE 5

Miscellanées

Afin d'être exhaustifs, il nous faut dire un mot d'autres articles, étrangers aux précédents chapitres. Ces publications entrent dans le cadre de thèses postérieures à celles de Chevrot et Khelil (cf. Chapitre 1), sauf [**P2**].

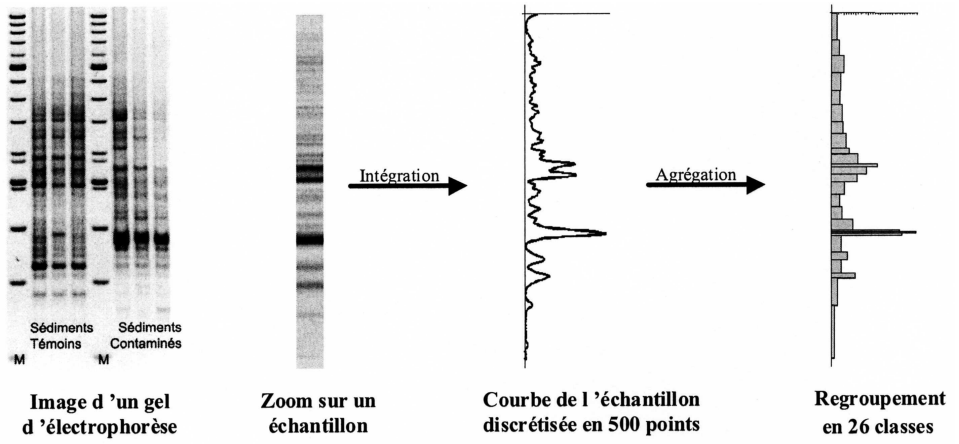
Nous allons procéder dans l'ordre chronologique, en commençant par [**P2**], portant sur la segmentation des profils d'éboulis alpins. Chaque profil était décrit par un ensemble de mesures de la pente du versant en fonction de la distance au sommet de l'éboulis. Notre but était de le diviser en plusieurs parties approximativement linéaires, associées à différents processus modelant le relief. Pour ce, nous avons proposé d'ajuster les profils par des splines cubiques d'approximation. Les points de rupture recherchés étaient assimilés aux plus fortes discontinuités de la dérivée seconde de la spline d'approximation de chaque talus, c'est-à-dire aux points de courbure maximale du relief.

Nous avons plus tard été co-auteur de trois publications attachées à la thèse D. Nerini [**Ne00**]. Son sujet était en grande partie centré sur l'utilisation d'arbres de classification et de régression (CART) en Océanologie. Un des défauts de cette famille de méthodes est leur manque de robustesse : un changement mineur de l'échantillon d'apprentissage peut modifier de manière importante l'arbre de décision. Pour améliorer les prédictions, nous avons proposé une procédure de transfert, valable aussi bien pour les arbres de régression [**P9**] que ceux de classification [**P10**]. Il s'agissait de renforcer l'homogénéité des feuilles de l'arbre en réaffectant (si nécessaire, si possible) les observations extrêmes de chaque feuille. Pour déterminer ces extrêmes, nous avons utilisé une variante robuste de la distance de Mahalanobis [**RL87**], ainsi que la valeur critique de cette distance recommandée par Penny [**Pe96**] pour les petits échantillons. Dans les deux cas, la précision de la prédiction a été améliorée d'environ 15%.

Notre troisième travail en commun [**P11**] était pour l'essentiel une application de l'Analyse Factorielle Multiple (AFM) [**EP90**]. Les données consistaient en 479 "journées" d'observation des principales variables physico-chimiques de l'Etang de Berre, mesurées à une station située à 8 mètres de fond. Chaque "journée" recensait les valeurs horaires de la température, de la salinité et du pourcentage d'oxygène dissous. Les trois chroniques journalières ont été au préalable décomposées en séries de Fourier discrètes, ce qui présente l'avantage (il y en a d'autres : voir Chapitre 1) d'isoler la première harmonique de chaque signal (sa moyenne quotidienne). En effet, quand on traite ce genre de données, on observe généralement que la variance de la moyenne quotidienne correspond à une énorme part de la variance totale, que nous pouvons ici immédiatement séparer du reste. Nous avons ensuite analysé l'ensemble des journées au moyen d'une AFM. Cette méthode, utilisant une structuration *a priori* des variables en groupes, nous a permis de tenir compte de l'hétérogénéité des variables. Nous avons ainsi constitué le groupe des trois moyennes quotidiennes, et un groupe pour l'ensemble des autres harmoniques associées à chacune des variables physico-chimiques (température, salinité, oxygène), car elles s'expriment dans des unités différentes. En pondérant de manière équilibrée chacun de ces quatre groupes (le "poids" d'un groupe est sa dimension), l'AFM nous a permis de classer l'ensemble des journées en cinq états types quotidiens qui résumaient bien le degré de stratification ou de mélange des eaux de l'étang.

La thèse de Gilles Miralles [**Mi07**] nous a ensuite donné l'occasion de nous mêler d'Ecologie Microbienne. Dans cette thèse, dont le sujet était "l'étude de l'impact d'une pollution par des hydrocarbures sur les communautés bactériennes", l'auteur a utilisé la technique d'empreinte moléculaire RISA (Ribosomal Intergenic Spacer Analysis). Cette méthode sépare les fragments d'ADN (baignant dans un gel conducteur soumis à une différence de potentiel) suivant leur taille,

FIG. 5.1. ***Quelques étapes du traitement d'un gel.***



les plus petits fragments migrant le plus rapidement vers la borne positive du dispositif. Quand la migration est achevée, le gel est coloré au moyen d'un réactif, et l'image obtenue est verticalement intégrée par le logiciel d'acquisition. La structure génétique de toute la communauté bactérienne est finalement résumée par ce profil, qui est très finement échantillonné (plusieurs centaines de "bandes"). Ces profils ne sont pas aisément interprétables, du fait du nombre important de bandes, et du grand nombre de discontinuités qu'ils présentent. Nous avons donc proposé une méthode permettant de simplifier ces profils. Elle consiste en bref à agréger progressivement les bandes contiguës, de façon à simultanément conserver au mieux la structure du profil moyen (caractérisé par sa néguentropie) et celle du nuage des profils, résumé par ses k (fixé) premières composantes principales. Pour parvenir à ce deuxième objectif on a utilisé à nouveau (voir Section 4) la distance de Krzanowski, aussi bien dans le cadre de l'ACP [**Kr87**] que dans celui de l'AFC [**Kr93**].

Un exemple de traitement est montré sur la Figure 5.1, où l'on peut voir comment un profil est simplifié au cours de l'analyse. La programmation en R de la méthode a été assurée par F. Bremond, dans le cadre de son stage de DESS [**Br04**], et une première application a été récemment publiée [**Mi07**]. Un article méthodologique à ce sujet est en gestation.

Enfin, le sujet de la thèse (en cours) de Bastien Mérigot est l"Analyse multi-composantes de la diversité spécifique marine". La biodiversité est un concept assez flou, et tellement multiforme que certains en sont venus à le trouver insignifiant [**Hu71**]. Il est donc salutaire d'essayer de se repérer dans la jungle des indices en examinant conjointement leurs propriétés mathématiques (concavité, par exemple), statistiques et écologiques. Dans [**P19**], les relations entre une douzaine d'indices de diversité courants sont étudiées au moyen du coefficient de corrélation de rangs de Spearman. L'ACP de la matrice de corrélations associée a permis le regroupement de ces indices en six composantes complémentaires pour décrire la diversité des données. Dans un deuxième article [**P20**], il est montré sur un autre jeu de données que les mêmes indices s'agrègent encore en six groupes pertinents, très peu différents de ceux de [**P19**]. De plus, les relations entre indices étaient semblables pour les différentes strates bathymétriques étudiées. On a observé finalement que l'interprétation écologique des indices elle-même pose parfois question : des indices espérés complémentaires se sont révélés redondants dans les deux études.

Perspectives

Je terminerai ce document par une liste de travaux en cours (ou sur le point de l'être), successivement dans l'ordre des quatre chapitres.

La cytométrie en flux : mise sur orbite

Parmi les méthodes d'investigation à l'échelle cellulaire, la cytométrie en flux est une technologie qui s'impose de plus en plus en microbiologie marine car elle permet l'analyse à haute vitesse des cellules (plusieurs milliers de cellules analysées par seconde). Les cellules sont entraînées par un liquide vecteur et sont interceptées une par une par une source lumineuse (un ou plusieurs lasers). Les propriétés optiques de diffusion de la lumière et d'émission de fluorescence permettent ensuite de discriminer les différents groupes de cellules présents, ainsi que leur concentration dans l'échantillon.

Notre laboratoire est équipé d'un nouveau type de cytomètre en flux, le CytoSub, conçu spécialement pour l'analyse du phytoplancton. Sa conception permet l'analyse de grosses particules et l'enregistrement du profil des signaux générés lors de l'interception des particules par le faisceau laser. Chaque cellule est donc associée à plusieurs "canaux" correspondant à la mesure de différents paramètres (taille, texture (présence d'organelles), fluorescence à différentes longueurs d'onde).

Notre intention est d'identifier les portions de ce signal multivarié (extraites automatiquement par l'appareil), c'est-à-dire de les associer à telle ou telle espèce. C'était exactement le but affiché dans [**P7**], où a été montré l'intérêt conjoint de la transformée de Fourier et des distances orbitales pour résoudre ce genre de problèmes. Dans ce travail, nous devions aussi extraire les portions d'intérêt, mais nous aurons ici d'autre difficultés pratiques. D'abord, contrairement aux bulles d'air générées par A. Khelil, les cellules phytoplanctoniques ne sont ni sphériques ni calibrées (même dans le cas monospécifique). Ensuite, l'abondance des données sera une source de difficultés : si des milliers (voire des millions) de cellules sont analysés, comment traiter la gigantesque table d'interdistances associée (rappelons que les distances orbitales ne sont *a priori* pas euclidiennes) ?

Ce projet, mené en collaboration avec des collègues du LMGEM (M. Denis, G. Gregori, D. Nerini, M. Thyssen, A.F. Yao), fait l'objet d'une demande de thèse, et d'une demande de projet ANR (PHYRMED).

L'avenir de la Rareté

Norbert Wiener, parlant de l'épuisement des ressources naturelles dans un ouvrage visionnaire [**Wi71**] (paru en 1952, et scandaleusement introuvable aujourd'hui) évoquait le Goûter Fou de "Alice au Pays des Merveilles" :

> *Quand le thé et les gâteaux étaient épuisés devant une chaise, le Chapelier Fou et le Lièvre de Mars ne trouvèrent rien de plus naturel que de "circuler" et d'occuper la chaise suivante. Quand Alice demanda ce qui se produirait lorsqu'ils seraient revenus à leurs places initiales, le Lièvre de Mars changea de conversation.*

Il en est de même actuellement pour beaucoup d'espèces sauvages, qui "circulent" à la surface de la planète, sous l'effet du Désordre Climatique ! Notre petite théorie de la rareté a suscité l'intérêt de nombre d'écologues, qui se sont néanmoins généralement abstenus de la pratiquer, peut-être pour ne pas en altérer la pureté, ou plus probablement à cause de l'aridité de certains articles [**P12*, P13**] et du choix délicat de α_0. Cependant, par les temps qui courent, il ne serait

pas indifférent de savoir si telle espèce se fait rare ici (et apparaît là, ou disparaît), alors que d'autres, naguère rares, pullulent. Cela m'incite donc à un certain optimisme : cette théorie étant loin d'avoir épuisé ses potentialités, elle est sûrement pleine d'avenir !

L'ACP de mesures

Mon travail en cours sur ce sujet gravite autour du Corollaire 1.2 du Chapitre 4, qui est d'une grande importance pratique : il démontre que l'analyse deviendrait très simple si les données étaient échantillonnées en des fractiles de la distribution de référence μ. J'ai donc entrepris, pour une distribution de référence empirique donnée (une granulométrie particulière, par exemple) de déterminer un système "optimal" de fractiles, c'est-à-dire proche (au sens de la distance de Hausdorff) d'une partie (de même cardinal) des noeuds de la grille imposée par l'appareillage. Il est donc nécessaire de localiser les fractiles de μ en interpolant la f.r. associée. J'utilise pour cela l'opérateur de Bernstein, qui possède d'excellentes propriétés statistiques **[BC02]**, ainsi que ses itérées, bien supérieures dans la convergence numérique **[Sa04]**.

Cette méthode est actuellement appliquée à un problème de bioturbation, en collaboration avec des collègues du LMGEM (E. Duport, G. Stora et A.F. Yao) et F. Gilbert (EcoLab, UMR 5245, Toulouse). Concrètement, il s'agit de mettre en évidence d'éventuelles modifications de la granulométrie d'un sédiment, imputables à la bioturbation par des organismes macrobenthiques.

Ce travail se déroule dans le cadre d'un programme transverse du LMGEM, qui sera peut-être prolongé par un projet ANR..

Le mouvement Brownien fractionnaire et le plancton

Dans le Chapitre 4, nous avons proposé de simuler l'environnement turbulent du phytoplancton par transformation d'un bruit Gaussien fractionnaire **persistant** ($H > 0.5$). D'une manière harmonieuse mais surprenante, nous avons mis en évidence dans plusieurs chroniques de comptages de cellules phytoplanctoniques en culture, l'existence d'une composante aléatoire, qui semble être une réalisation d'un mouvement Brownien fractionnaire **antipersitant**. Or, Cioczec-Georges et Mandelbrot **[CM95]** on démontré qu'un tel processus peut être obtenu comme "somme fractale" de micro-impulsions. D'après cet article, notre composante pourrait résulter directement de la distribution de durée de vie des cellules. Mais il se pourrait aussi qu'elle soit liée au cycle cellulaire des organismes ou, plus trivialement, que ce soit un pur artefact induit par le système de commande du chémostat ! Ce point fera l'objet de futures recherches, menées en collaboration avec D. Nerini et A. Sciandra (LOV, Station marine de Villefranche-s.m.).

Les premiers résultats on été présentés à COMPSTAT2006 (C. Manté, D. Nerini, A. Sciandra & A.F. Yao, *Estimating the Hurst exponent of self-similar processes. Applications in Marine Ecology*, Rome, 28 Août-1 Septembre 2006).

Bibliographie

[P1] C. Manté, *Analyse en Composantes Principales d'un processus multiple non stationnaire : Une application à des données météorologiques*, Statistique et Analyse des Données, 14, 2, 25-53, 1989.

[P2] B. Francou and C. Manté, *Analysis of the Segmentation in the Profil of Alpine Talus Slope*, Permafrost and Periglacial Processes,1 , 1 , 53-60, 1990.

[P3] P. Chevrot, C. Manté et B.A. Thomassin, *Sclérochronologie de Porites lutea (Sclératinaire hermatypique) à Mayotte (SW océan indien) : Nouvelle approche par l'Analyse en Composantes Principales dans le domaine des fréquences*, C.R. Acad. Sci. Paris, t.318, série II, 803-808, 1994.

[P4] C. Manté, J.C. Dauvin and J.P. Durbec, *Statistical method for selecting representative species in multivariate analysis of long-term changes of marine communities. Applications to macrobenthic communities from the Bay of Morlaix*, Marine Ecology Progress Series, 120, 243-250, 1995.

[P5] P. Chevrot et C. Manté, *Une nouvelle méthode d'analyse statistique de radiographies de coupes de coraux. Détermination des bandes annuelles de croissance*, J. Rech. Océanogr., 20, 79-83, 1996.

[P6] C. Manté, J.P. Durbec et J.C. Dauvin, *Analyse de l'évolution temporelle de communautés macrobenthiques à partir des probabilités de présence des espèces*, Oceanologica Acta, 20, 1, 71-79, 1997.

[P7] A. Khelil, C. Manté et P.M. David, *Localisation et discrimination de signaux acoustiques de bulles d'air par des techniques statistiques*, Traitement du Signal, 14, 2, 151-159, 1997.

[P8*] C. Manté, *The use of regularization methods in computing Radon- Nikodym derivatives. Application to grain-size distributions*, SIAM Journal on Scientific Computing, 21, 2, 455-472, 1999.

[P9] D. Nérini, J.P. Durbec and C. Manté, *Analysis of oxygen rate time series in a strongly polluted lagoon using a regression tree methods*, Ecological Modelling, 133, 95-105, 2000.

[P10] D. Nérini, J.P. Durbec, C. Manté, F. Garcia and B. Ghattas, *Forecasting physicochemical variables by a classification tree method. Application to the Berre lagoon (South France)*, Acta Biotheoretica, 48, 181-196, 2000.

[P11] D. Nérini, C. Manté, J.P. Durbec et F. Garcia, *Une méthode statistique de détermination de séquences caractéristiques dans une série temporelle de plusieurs variables. Application à la physico-chimie des eaux de l'étang de Berre,* C.R. Acad. Sci. Paris, Sciences de la Terre et des Planètes, 332, 457-464, 2001.

[P12*] C. Manté, B. Elkaim and J.C. Dauvin, *Methods for selecting dominant species in ecological series. Application to marine macrobenthic communities from the English Channel*, J. Rech. Océanogr., 26, n° 1-2, 29-36, 2001.

[P13] C. Manté, J. Claudet and C. Rebzani-Zahaf, *Fairly processing rare and common species in Multivariate Analysis of ecological series. Application to communities from Algiers harbour*, Acta Biotheoretica, 51, 4, 277-294, 2003.

[P14] C. Manté, J.P. Durbec and J.C. Dauvin, *A functional Data-Analytic approach to the classification of species according to their spatial dispersion. Application to a marine macrobenthic community from the Bay of Morlaix (western English Channel)*, Journal of Applied Statistics 32, 8, 831-840, 2005.

[P15*] C. Manté, A.F. Yao and C. Degiovanni, *Principal Components Analysis of measures, with special emphasis on grain-size curves*, Computational Statistics & Data Analysis, 51, 4969-4983, 2007.

[P16*] C. Manté, *Application of resampling and linear spline methods to spectral and dispersional analyses of long-memory processes*, Computational Statistics & Data Analysis, 51, 4308-4323, 2007.

[P17] G. Miralles, D. Nérini, C. Manté, M. Acquaviva, P. Doumenq, V. Michotey, S. Nazaret, J. C. Bertrand and P. Cuny, *Effects of spilled oil on bacterial communities of Mediterranean coastal anoxic sediments chronically subjected to oil hydrocarbon contamination*, Microbial Ecology, 54, 646-661, 2007.

[P18] J.C. Poggiale, P. Auger, D. Nerini, C. Manté and F. Gilbert, *Global production increased by spatial heterogeneity in a population dynamics model*, Acta Biotheoretica, 53, 359-370, 2005.

[P19] B. Mérigot, J. A. Bertrand, N. Mazouni, C. Manté, J. P. Durbec and J. C. Gaertner, *A multi-component analysis of species diversity of groundfish assemblages on the continental shelf of the Gulf of Lions (north-western Mediterranean Sea*, Estuarine, Coastal and Shelf Science, 73, vol. 1-2, 123-136, 2007.

[P20] B. Mérigot, J.A. Bertrand, J.C. Gaertner, J.P. Durbec, N. Mazouni and C. Manté, *The multi-component structuration of the species diversity of groundfish assemblages of the east coast of Corsica (Mediterranean sea) : variation according to the bathymeric strata*, Fisheries Research, 2007, sous presse.

[AE01] M. Arnaud, X. Emery, C. de Fouquet, M. Brouwers et M. Fortier, *L'analyse krigeante pour le classement d'observations spatiales et multivariées*, Revue de Statistique Appliquée, XLIX,2, 45-67, 2001.

[BC02] G.J. Babu, A.J. Canty, Y.P. Chaubey, *Application of Bernstein polynomials for smooth estimation of a distribution and density function*, Journal of Statistical Planning and Inference, 105, 377-392, 2002.

[BL03] J.M. Bardet, G. Lang, G. Oppenheim, A. Philippe, M.S. Taqqu, *Generators of Long-Range dependent processes : a survey*, in : P. Doukhan, G. Oppenheim, M.S. Taqqu (Eds.), Theory and applications of long-range dependence, Birkhauser, Boston, pp. 557-577, 2003.

[Ben99] R. Benchley, *La vie périlleuse du chanteur de basse* (ed.bilingue), Editions du Rocher, Monaco, 1999.

[Ber94] J. Beran, *Statistics for long-memory processes*, Chapman & Hall, London, 1994.

[BL87] D. Bosq et J.P. Lecoutre, *Théorie de l'estimation fonctionnelle*, Economica, Paris, 1987.

[Bo99] P. Bogaert, *On the optimal estimation of the cumulative distribution function in presence of spatial dependence*, Mathematical Geology, 3, 2, 213-239, 1999.

[Br04] F. Bremond, *Codage optimal pour l'analyse multivariée de profils d'électrophorèse et de spectres de taille de zooplancton*, Stage de DESS de l'Université Blaise Pascal - Clermont-Ferrand, 2004.

[CP76] F. Caillez et J.P. Pagès, *Introduction à l'Analyse des Données*, SMASH, Paris, 1976.

[CH95] P. Chevrot, *Analyse statistique des images. Application à des séries d'images en Océanologie*, Thèse de Doctorat de l'Université de la Méditerranée, Spécialité : Océanologie, 1995.

[CM95] R. Cioczec-Georges and B.B. Mandelbrot, *A class of micropulses and antipersistent fractional Brownian motion*, Stochastic Processes and their Applications, 60, 1-18, 1995.

[DM93] G. Der Mégréditchian, *Le traitement statistique des données multidimensionnelles. Application à la Météorologie*, Cours et manuels N°6, Météo-France, 1993.

[De74] J.C. Deville, *Méthodes statistiques et numériques de l'analyse harmonique*, Annales de l'INSEE, 15, 1974.

[DN97] C.R. Dietrich and G.N. Newsam, *Fast and exact simulation of stationary gaussian processes through circulant embedding of the covariance matrix*, SIAM Journal on Scientific Computing, 18, 4, 1088-1107, 1997.

[Du06] D. Dubois, *Possibility theory and statistical reasoning*, Computational Statistics & Data Analysis, 51, 47-69, 2006.

[EP90] B. Escofier et J. P. Pagès, *Analyse factorielles simples et multiples*, Dunod, Paris, 1990.

[EV06] G. Estévez and P. Vieu, *Nonparametric estimation under long memory dependence*, Journal of Nonparametric statistics, 15, 4-5, 535-551, 2003.

[FV06] F. Ferraty and Ph. Vieu, *Nonparametric Functional Data Analysis. Theory and practice,* Springer series in Statistics, Berlin, 2006.

[Fr02] U. Frisch, *La turbulence*, L'Université de tous les savoirs, vol. 13, Odile Jacob, Paris, 159-172, 2002.

[Ga94] K.J. Gaston, *Rarity*, Population and community biology series, 13, Chapman & Hall, London, 1994.

[GK97] *The biology of rarity. Causes and consequences of rare-common differences*, W.E. Kunin and K.J. Gaston eds, Chapman & Hall, London, 30-47, 1997.

[Gaus35] G. F. Gause, *Vérifications expérimentales de la théorie mathématique de la lutte pour la vie*, Actual. scient. indust, 277, 1-62, 1935.

[Gaut82] W. Gautschi, *On generating orthogonal polynomials*, SIAM Journal on Scientific Computing, 3, 3, 289-317, 1982.

[GL96] G. H. Golub and F.V.L. van Loan, *Matrix computation (third edition)*, The John Hopkins University Press, Baltimore and London, 1996.

[GGY89] F. Gourd, J.P. Gauthier et H. Younes, *Une méthode d'invariants de l'analyse harmonique en reconnaissance des formes*, Traitement du Signal, 6, 3, 161-178, 1989.

[GN03] M. Gradinaru and I. Nourdin, *Approximation at first and second order of m-order integrals of the fractional Brownian motion and of certain semimartingales*, Electronic Journal of Probabilities, 8, 1-26, 2003.

[Han98] P. C. Hansen, *Rank-deficient and discrete ill-posed problems*, SIAM Monographs on Mathematical modeling and Computation, 1998.

[Ha60] G. Hardin, *The competitive exclusion principle*, Science, 131, 1292-1297, 1960.

[Hu71] S. H. Hurlbert, *The nonconcept of species diversity : a critique and alternative parameters*, Ecology, 52, 577-586, 1971.

[Ka76] Y. Katznelson, *An introduction to Harmonic Analysis*, Dover Publications, New York, 1976.

[Ke89] D. G. Kendall, *A survey of the Statistical Theory of Shape (with discussion)*, Statistical Science, 4, 2, 87-120, 1989.

[AK95] A. Khelil, *Acoustique de bulles d'air dans une colonne d'eau. Identification et discrimination*, Thèse de Doctorat de l'Université de la Méditerranée, Spécialité : Océanographie physique, 1995.

[Kl1872] F. C. Klein, *Le programme d'Erlangen, Considérations comparatives sur les recherches géométriques modernes*, Ed. Jacques Gabay, Paris, 1991.

[Kru34] W. C. Krumbein, *Size frequency distribution of sediments*, Journal of Sedimentary Petrology, 4, 2, 65-77, 1934

[Kr87] W. J. Krzanowski, *Selection of variables to preserve Multivariate Data Structure, using Principal Components*, Applied Statistics, 36, 22-33, 1987.

[Kr93] W. J. Krzanowski, *Attribute selection in correspondence analysis of incidence matrices*, Applied Statistics, 42, 3, 529-541, 1993.

[KB07] W. J. Krzanowski and T. C. Bailey, *Extraction of spatial features using factor methods illustrated on stream sediment data*, Mathematical Geology, 39, 1, 69-85, 2007..

[Lag06] Y. Lagadeuc, *Nutrient fluxes toward phytoplankton : is it useful to consider turbulence intermittency ?*, Acta Biotheoretica, 53, 371-379, 2006.

[Lar92] H.J. Larson, *Least square estimation of linear splines with unknown knot location*, Computational Statistics & Data Analysis, 13, 1-8, 1992.

[Le55] E. L. Lehmann, *Ordered families of distributions*, Annals of Mathematical Statistics, 26, 399-419, 1955.

[LK00] H. Le and A. Kume, *Detection of shape changes in biological features*, Journal of Microscopy, 200, 2, 140-147, 2000.

[LS07] S. Lovejoy and D. Schertzer, *Scaling and multifractal fields in the solid earth and topography*, Nonlinear Processes in Geophysics, 14, 465-502, 2007.

[LM78] P. Lütz et D. Maïti, *Classification automatique d'après la distance entre orbites : application à la physique corpusculaire*, Les Cahiers de l'Analyse des Données, 3, 4, 449-458, 1978.

[Ma07] A. E. Magurran, *Species abundance distributions over time*, Ecology Letters, 10, 347-354, 2007.

[MH03] A. E. Magurran & P. A. Henderson, *Explaining the excess of rare species in natural species abundance distributions*, Nature, 422, 714-716, 2003.

[Man89] B. B. Mandelbrot, *Les objets fractals*, Nouvelle Bibliothèque Scientifique Flammarion, Paris, 3ème ed., 1989.

[Man97] B. B. Mandelbrot, *Fractales, hasard et finance*, Flammarion, Paris, 1997.

[Mi07] G. Miralles, *Devenir d'une contamination pétrolière dans des sédiments côtiers infralittoraux et son impact sur les communautés bactériennes*, Thèse de Doctorat de l'Université de la Méditerranée, Spécialité : Biosciences de l'Environnement, Ecologie Bactérienne, 2007.

[Ne00] D. Nérini, *Analyse statistique de processus physiques et chimiques en Océanologie Côtière à l'aide d'une méthode de régression et de classification par arbre décisionnel. Application à l'étude d'un milieu fortement perturbé : l'Etang de Berre*, Thèse de Doctorat de l'Université de la Méditerranée, Spécialité : Sciences de l'environnement marin, 2000.

[No01] J. P. Nolan, *Maximum likelihood estimation and diagnostics for stable distributions*, In : O. E. Barndorf-Nielsen, T. Mikosch and S. I. Resnick (Eds.), *Lévy Processes : Theory and Applications*, Birkäuser, Boston, 379-400, 2001.

[NB05] E. Nowak and A. Bar-Hen, *Influence function and Correspondence analysis*, Journal of Statistical Planning and Inference, 134, 26-35, 2005.

[Pe96] K.I. Penny, *Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance*, Applied Statistics, 45, 1, 73-81, 1996.

[Po43] H. Poincaré, *La Science et l'Hypothèse*, Flammarion, Paris, 1943.

[RS05] J.O. Ramsay and B.W. Silverman, *Functional Data Analysis*, Springer series in Statistics, New York, 2005.

[RL87] P.J. Rousseeuw and A.M. Leroy, *Robust regression and Outlier Detection*, John Wiley & Sons, New York, 1987.

[Sa04] A. Sahai, *An iterative algorithm for improved approximation by Bernstein's operator using statistical perspective*, Applied Mathematics and Computation, 149, 327-335, 2004.

[ST94] G. Samorodnitsky and M. S. Taqqu, *Stable non-Gaussian random processes*, Chapman & Hall, London, 1994.

[SL02] D. Schertzer, S. Lovejoy and P. Hubert, *An introduction to stochastic multifractal fields*, Mathematical Problems in Environmental Science and Engineering, A. Ern and Liu Weiping (eds.), Series in Contemporary Applied Mathematics, vol.4, Higher Education Press, Beijing, p. 106-179, 2002.

[SSL01] L. Seuront, F. Schmitt and Y. Lagadeuc, *Turbulence intermittency, small-scale phytoplankton patchiness and encounter rates in plankton : where do we go from here ?*, Deep-Sea Research I, 48, 1199-1215, 2001.

[SSN02] E. Sicard, R. Sabatier, H. Niel and E. Cadier, *A new approach in space-time analysis of multivariate hydrological data : application to Brazil's Nordeste region rainfall*, Water Resources Research, 38, 12, 1319, 55, 1-10, 2002.

[Si86] B. Silverman, *Density Estimation for Statistics and Data Analysis*, Monographs on Statistics and Applied Probability N°29, Chapman & Hall, London, 1986.

[SW95] H.L. Smith and P. Waltman, The theory of chemostat, Cambridge University Press, Cambridge, 1995.

[TL00] I. Tchiguirinskaia, S. Lu, F.J. Molz, T.M. Williams and D. Lavallée, *Multifractal versus monofractal analysis of wetland topography*, Stochastic Environmental Research and Risk Assesment, 14, 8-32, 2000.

[Th93] R. Thom, *Prédire n'est pas expliquer*, Flammarion, Paris, 1993.

[UB73] J. Ulmo et J. Bernier, *Eléments de décision statistique*, PUF, Paris, 1973.

[UO04] W. Ulrich and M. Ollik, *Frequent and occasional species and the shape of relative-abundance distributions*, Diversity and Distributions, 10, 263-269, 2004.

[Wa98] H. Wackernagel, *Multivariate Geostatistics*, Springer Verlag, Berlin, 1998.

[Wi71] N. Wiener, *Cybernétique et société*, Union Générale d'Editions, Collection 10/18, 1952, rééd. 1971.